

Stig Nordbeck
Bengt Rystedt

Computer cartography

1973

STUDENTLITTERATUR





Digitized by the Internet Archive
in 2026

Computer
cartography

Computer cartography

the mapping process revisited

by David H. Johnston

Stig Nordbeck

Bengt Rystedt

Computer cartography

The mapping system NORMAP

Location models

Distribution

Studentlitteratur Lund



© Stig Nordbeck Bengt Rystedt 1972

Printed in Sweden

Carl Bloms Boktryckeri AB
Lund 1972

ISBN 91-44-04651-0

Contents

9	FOREWORD
11	INTRODUCTION
21	PART I. THE MAPPING SYSTEM NORMAP
23	1 COMPUTERIZED MAP CONSTRUCTION
24	1.1 SYMAP
24	1.2 LINMAP
25	1.3 NORMAP
26	1.4 NORMAP's programming languages
27	2 TERMS, CONCEPTS AND NOTATION
27	2.1 Function
27	2.2 Program, procedure and subroutine
28	2.3 Constant, parameter and variable
28	2.4 Coordinates
29	2.5 Grid point and grid net
30	2.6 Data point
30	2.7 Reference interval and reference area
30	2.8 Equidistance
31	2.9 Overlapping
32	3 SQUARE NET MAPS AND THE SQUARE GRID NET
32	3.1 The square net map program NORK
34	3.2 The overlapping program NORKO
35	3.3 The MAPOUT program
36	3.4 The MAPIN program
37	4 REGULAR TRIANGULAR GRID NETS AND ISARITHMIC MAPS
37	4.1 Maps and 3-dimensional diagram
38	4.2 Polygon diagram and isarithmic maps
38	4.3 Location of the isarithms by interpolation
40	4.4 The joining up of line segments to form isarithms
40	4.5 The NORI program
40	4.6 The interpolation program NORIP
43	4.7 The transformation routine MAPTRF
43	4.8 The storage of numerical maps on magnetic tape
44	5 IRREGULAR GRID NETS
45	5.1 Approximate calculation of values at grid points
48	5.2 Storage of irregular grid nets
48	5.3 The NORIPO program for irregular grid nets
51	5.4 Transformation of a regular net into an irregular net
51	5.5 The graphical output procedure NORIPL
53	6 THE CALCULATION OF VOLUMES AND AREAS
53	6.1 Polygon and the determinant formula
55	6.2 The calculation of volumes using Simpson's formula

57	6.3	Volume calculations — a numerical example
59	6.4	The NORVSI program for volume calculations
59	6.5	Numerical maps and the calculation of volumes
60	6.6	Numerical terrain models
63	7	THE FITTING OF A POLYNOMIAL TO A SET OF POINTS
63	7.1	The fitting of a plane to a set of points in space
64	7.2	Truncation errors and the mean transformation
64	7.3	The coefficient table, the general case
65	7.4	Logarithmic transformation
66	7.5	The solving of equation systems
67	7.6	Correlation calculations
67	7.7	The curve fitting program NORMKA
69	7.8	Trend surface analysis
69	7.9	Coordinate transformation
71	8	COMPARISON OF ISARITHMIC MAPS
71	8.1	Isarithm overlay
73	8.2	The quotient and difference methods
73	8.3	The residual method
78	8.4	The correlation method
79	8.5	Auto-correlation
82	9	MAPPING OF TRAFFIC FLOWS
82	9.1	Straight line journeys
83	9.2	Linear journeys and the NORILK procedure
83	9.3	Journeys via actual road networks and the NORIRK procedure
86	10	RANGE MAPS
86	10.1	Constant reference volume, dot maps
87	10.2	Range calculation algorithms
88	10.3	The reference area's shape
90	10.4	Range maps of Skåne
91	10.5	Fraction maps
94	10.6	Transport cost maps
96	10.7	The NORREM program
96	10.8	Isarithmic population maps and the location of facilities
98	11	POPULATION POTENTIALS
99	11.1	Weights other than distance
100	11.2	The contribution made by place PI to the potential
102	11.3	Square net maps and the procedure NORPOT
103	12	LOCATION CALCULATIONS
103	12.1	Location of one facility
104	12.2	Total transportation distance and potentials
105	12.3	Location of one facility, iteration method
106	12.4	Location of several facilities, iteration method
107	12.5	The transportation cost function COST
108	12.6	The NORLOC program
109	13	THE SHORTEST DISTANCE BETWEEN TWO NODES IN A ROAD NETWORK
109	13.1	The cascade method and the NORCAS procedure
114	13.2	The ellipse method and the procedure NORAN

118	14 BARRIERS
118	14.1 Range maps, barriers and location calculations
124	14.2 Straight line barriers and the NORLAS program
126	14.3 Polygon barriers and the program NORPAS
126	14.4 Combinations of several barriers and NORAS
131	15 SORTING ON THE BASIS OF POLYGONS
131	15.1 Tests for convexity in polygons
132	15.2 Convex polygons and NORPCO
133	15.3 The orientation theorem for concave polygons and NORP
134	16 DIRECTIONAL INFORMATION AND FABRIC DIAGRAMS
135	16.1 The description of the direction of a plane in space
136	16.2 The description of the direction of lines in space
137	16.3 Fabric diagram — planes
138	16.4 Fabric diagram — lines
139	16.5 The projection of the half sphere on to an xy plane
142	16.6 The projection on to the xy plane of the point of intersection between a line (b -axis) and the surface of the half sphere
142	16.7 Values at the periphery of the net
146	16.8 The fabric diagram program NORIGE
147	16.9 Area fabric diagrams and NORYG
148	16.10 The drawing of the fabric diagram and NORGEF
150	17 THE MAIN PROGRAM OF THE NORMAP SYSTEM
150	17.1 Data
152	17.2 The main program ISPLOT
155	PART II. LOCATION MODELS
157	1 THE SPATIAL PATTERN OF CENTRAL PLACES
158	1.1 Central place theories
160	1.2 The hierarchy of central places and the rank size rule
161	1.3 Criticisms of the basic assumptions
163	2 SPONTANEOUS LOCATION
164	2.1 Business activity in Örkened at the end of the 19th century
166	2.2 The quarry workers' dwellings
167	2.3 Service facilities in Örkened 1920 and 1970
172	2.4 The development of central places and the spontaneous location of service facilities
174	3 THE CENTRAL PLACE PATTERN IN SOUTHERN SWEDEN
174	3.1 The distribution of population in southern Sweden 1965
175	3.2 The size of the reference area and the net's equidistance
176	3.3 The testing of the central place pattern
182	4 THE DISTRIBUTION OF POPULATION IN SWEDEN IN THE YEARS 1855, 1917 AND 1965
182	4.1 The railway investigation committee's map
186	4.2 Thure von Mentzer's population dot map
187	4.3 Sten de Geer's population dot map
192	4.4 The distribution of population in Sweden in 1965
193	4.5 The accuracy of the different maps

194	5 A MATHEMATICAL LOCATION MODEL
194	5.1 The multiple location problem
194	5.2 Locating one single activity
196	5.3 Locating several facilities in a 1-dimensional population distribution
198	5.4 Locating two facilities in a 2-dimensional population distribution
200	5.5 The general problem of locating multiple facilities
200	5.6 The multiple location algorithm
203	6 THE LOCATION FUNCTION AND LOCAL MINIMA
203	6.1 1-dimensional population distribution and local minima
205	6.2 Local minima and 2-dimensional population distribution
210	6.3 Local minima and groups of the population distribution
211	6.4 Practical planning problems and local minima
215	7 SENSITIVITY ANALYSIS OF THE LOCATION MODEL
217	7.1 Sensitivity to error in the data
223	7.2 Sensitivity to changes in the recording grid
226	7.3 The transportation cost function
241	7.4 Sensitivity to the use of straight line distance
247	7.5 Conclusions
248	8 THE STABILITY OF THE LOCATION MODEL WITH RESPECT TO CHANGES IN THE DISTRIBUTION OF POPULATION THROUGH TIME
248	8.1 Changes occurring within regions and between regions
250	8.2 The location of three facilities
251	8.3 The location of 5 facilities
255	8.4 The location of 10 facilities
256	8.5 The location of 15 facilities
256	8.6 The location of 23 facilities
260	8.7 Conclusions
262	9 STEPWISE REGIONALIZATION
267	10 STEPWISE LOCALIZATION
267	10.1 A comparison between stepwise and optimum localization
274	10.2 The best of N—I facilities
277	10.3 Conclusions
278	11 SIMULTANEOUS HIERARCHICAL LOCALIZATION
279	11.1 Costs incurred in the transportation of raw materials
279	11.2 The simultaneous location of two types of facility using a single map of population distribution as input data
280	11.3 The simultaneous location of a number of types of facility using a number of maps as input data
283	12 THE LOCATION MODEL AND PLANNING AT THE REGIONAL AND LOCAL LEVELS
283	12.1 Data for Malmö
285	12.2 The Malmö experiment
287	12.3 Locating colleges within a region
293	SUMMARY
295	1 The mapping system NORMAP
300	2 Location models
306	LITERATURE

Foreword

This report contains some of the results obtained in connection with research into the coordinate mapping method and its practical application in the social sciences, which has been carried on at Lund University between 1967 and 1972. Important parts of this research have been financed by The National Swedish Council for Building Research in Stockholm.

It was originally planned to describe the computer mapping system, developed in connection with the research project, in this report. This was to have been done in the form of, (1) a detailed description of the method lying behind each computer program in the system, (2) a presentation of each program in an explicit form, (3) instructions describing how each program is used, together with test examples, and finally (4) a collection of examples showing the application of each program in connection with practical planning problems. This was to have been done for each of the about 50 programs making up the system. During the writing of the report we discovered that it was necessary to divide it into sections, also a number of changes were made in the arrangement of the material. The programs forming NORMAP have a life of their own. Changes are constantly being made in the computer installation's operative system, which means that the NORMAP programs must be regularly brought up to date. Errors in the programs are discovered and corrected as they occur. New programs are added, as well as new versions of existing programs. It was therefore considered best to combine points 2 and 3 above into a special forthcoming report, which could be used as a system manual.

The methods used in NORMAP do not require continuous development, as is the case with the programs themselves. As a result the methods have been brought together in this report, which therefore consists of two parts. Part I contains descriptions of the different methods, together with information about the NORMAP system. This part also contains a number of sections which could form part of the manual but at the same time they are helpful when attempting to understand the system. Part II deals with the application of the system, mainly in connection with localization

problems. These examples have provided a practical application for the majority of the NORMAP programs.

Stig Nordbeck has written chapters 1—8 in part I and chapters 1—8 in part II. Bengt Rystedt is responsible for chapters 9—17 in part I and chapters 9—12 in part II of this report.

The maps, diagrams and illustrations have been drawn by Mrs Klara Laszlo, Department of building function analysis, University of Lund. Translation of the manuscript into English has been carried out by arkitekt SAR, tekn. lic. Anthony Cole, Lund.

Professor Carin Boalt of the Department of building function analysis has followed the progress of the project. She has read the manuscript and brought to our notice a number of alterations and improvements, as have professors Olof Nordström and Gunnar Törnqvist of the Department of human and economic geography of the University of Lund.

To all of them, institutions and private persons who has helped us in our work with this report, we would like to extend our warmest thanks.

Lund, Sweden, November 1972

Stig Nordbeck

Bengt Rystedt

Introduction

1 The planner needs data

Before any planning activity can be carried out it is necessary to obtain a picture of the existing situation which is as complete as possible. In the case of physical and social planning the data required to build up such a picture can be obtained in at least three ways. The first way is to extract the information from existing public registers or databanks, such as the register of births, marriages and deaths, and the cadastre. The second way is to use information gained in conjunction with official public surveys, such as the population, housing, and farming censuses and the land use maps. The third way to obtain the necessary data is to carry out a special survey of the area in question. Such special surveys of an area can include the mapping of its topography, its soil characteristics, the nature of its underlying rocks or the way in which it is drained. A further example of a special survey is provided by the use of interview techniques in an attempt to discover the social, contact and behaviour patterns existing in a residential district scheduled for renewal.

Special surveys are, in most cases, a very expensive way of obtaining the necessary data. As a result the planner attempts, as far as possible, to obtain his data in the first two ways, from databanks or from official public surveys. In most cases the data obtained from a databank is considerably more up to date than the corresponding census data. The contents of databanks are, as far as possible, held current; an example of this being provided by the new Swedish databank containing real estate data. This databank is to be continually updated and will provide a rich source of information for the planner. The databank will be capable of providing data describing an existing situation at other points in the time than census years.

When the planner has obtained the necessary data it must be processed. Many advantages can be gained by carrying out the processing at the same time as the data is extracted from the register. The extraction and processing of information from a databank can be carried out using a computer program package developed especially for this purpose. A program package of this type is described in part I of this report, NORMAP (Nordbeck's and Rystedt's MAPping system). It contains programs cap-

able of extracting data from a databank. Other programs in the package can process the data obtained. The results are presented in the form of tables, diagrams and maps such as square net maps and isarithmic maps. The first stages of a planning process, as described above, are only concerned with the preparation of the information on which decisions can be made.

In the case of building activities on a virgin site it is now necessary to decide what is to be built. When that decision has been made detail planning can proceed i.e. basic information can be prepared regarding the location of different items of the physical structure. This information can then be applied, by the decision maker, when deciding the locations for streets, service routes, buildings and service facilities. In order to simplify the making of these decisions the planner requires access to a collection of location models. These models can be used to generate optimum locations for the different items of the physical structure, at the same time the relative efficiency of one solution as against another can be determined. Even such relatively simple location models are capable of aiding the work of the planner. Part II of this report describes a number of simple models for the facility location problem. In order to make the use of these locational models possible it is necessary to give the spatial distribution of the facility customers in the form of a square net map.

2 Swedish official registers

The Swedes are probably the world's most registered population. A large number of official registers exist, also the individual citizen has the responsibility for making sure that these registers contain correct information about himself. A person who moves residence must, at the time he moves, inform the registration authority who updates the register. For historical reasons the church authorities are responsible for the registration of persons. The correct term for this procedure is therefore that one is "entered into the church register as resident at a particular real estate unit". The actual register is called the "parish book".

The oldest Swedish parish books are from the 17th century (Friberg 1953), but it was only during the first decades of the 18th century that they were introduced in every Swedish parish. As a result one can follow the pattern of population through more than 250 years using information from these books (Historical statistics of Sweden, 1955).

Of an even older date than the parish books are the "mantalslängder" (population registration lists), which were originally taxation lists. The

oldest are from the 16th century (Friberg & Friberg 1971). About 150 years ago the parish ministers became involved in preparing these registration lists and only then did their contents become completely reliable. Registration occurred once a year on the 1st of November. This meant that the head of every family had to go and give the necessary information about the members of his household. This information was then checked against that contained in the parish book, and the result used to produce an accurate list. The registration lists were then used to produce the taxation lists and register of electors. One pays tax to the local authority or parish in which one is registered and it is there that one is entitled to vote.

The yearly registration was a simple form of census. Now, in the days of computers, the material contained in the parish book is considered of sufficiently high quality to be used to prepare the yearly population registration list without the necessity of checking it against information obtained from the head of every family. This is now carried out every 5th years in conjunction with the national census.

Another of the Swedish official registers is the real estate register. This register contains information about the property such as its name and the name of its owner. If one acquires a property one must register the fact within a fixed period. If this is not done the property can revert to its former owner. The property register and the individual properties described in it are very important components in the system for registration of persons. A person is always registered in the parish book or the registration lists as being resident at a particular property. The exceptions to this rule are a small number of persons without any known address or persons that are registered as being associated with the parish.

3 Coordinate based real estate data

The parish books or the registration lists can be used to give detailed information about the number of persons resident at a particular property. This can in principle be done for every one of the last 250 years. If the location of the property is determined by locating it on a map a very detailed and accurate population map can be prepared. The finding of the property on a map is a very complicated job, which should be done in such a way that the result can also be used on other occasions. The best way of describing the location of a property is the use of coordinates based on a suitable coordinate system. In the case of Sweden it was found that the coordinate system used by the land use map was suitable. The property register investigatory committee proposed that an official measurement of

property coordinates should be carried out and that the property register should be expanded to include all properties' coordinates (Fastighetsregistrering. SOU 1966: 63). This is now being carried out by Centralnämnden för fastighetsdata (the Central Bureau for Real Estate Data). The whole project should be completed during the 1970's.

The coordinates given in the real estate register are highly accurate. The units used are 10 metres in rural areas and 2 metres in built-up areas. During the interim period before the complete register is finished approximate coordinate based population information for a number of Swedish counties has been prepared and square net maps of population distribution produced manually (Claeson 1964, Olsson 1966). An accuracy of 1 km has been used in country areas. In built-up areas the population has not been split up between grid squares but has been given as a lump sum.

4 The application of coordinate information

The property coordinates can be used for sampling, sorting or for map construction (Nordbeck 1969). If the coordinates of a property are known the corresponding map sheet can be easily identified using the simple numerical relationship between the coordinates and the sheet numbering.

When the coordinate is used in conjunction with population data samples can be drawn having the same spatial distribution as the population as a whole. Also computers can be used to prepare population data for freely chosen polygons (Nordbeck & Rystedt 1967) and also to carry out large scale map constructions (Nordbeck & Rystedt 1972). Of all the applications mapping by aid of computers is the most important, seen from the point of view of the planner. A surveyor would probably consider the ease with which properties can be located and the relationship between the coordinates and map sheet as being very important, too.

5 Building function analysis and the NORMAP mapping system

Parallel with the activities of the property registration committee, in the coordinate assignment and the inclusion of coordinates in the property register, work has been carried on at the University of Lund on methods for the application of the coordinate information in map production. Computers have been used widely in this work and the resulting programs have now been combined to present the NORMAP mapping system.

NORMAP has been developed at the Department of Building Function Analysis of the Lund Institute of Technology during the last 5 years.

That one should be interested in automated mapping in conjunction with building function analysis is not as obvious as it is in the case of e.g. geography. Building function analysis does not concern itself only with how buildings function seen from the user's point of view, although of course this is the most important part of the subject. Research projects lying within this part of the subject, which are being carried on at the Department of building function analysis in Lund, include, "Adaptable dwellings" (Olsson & Nilsson 1970, Nilsson, Thorén & Åhlund 1971) and "Building programme methods" (Bergström 1971a). In most cases it is not sufficient to study how the individual building functions. The field must be widened to include how the building functions from the user's point of view in relation to its immediate surroundings. Research projects being carried on at the University of Lund which reflect this wider definition of the subject include, "The characteristics of the physical and social environment" (Hellberg & Lindberg 1969, Dahlgren, Hellberg & Lindberg 1971), "Service facilities in housing areas" (Pedersen 1971) and "Schoolchildren and traffic" (Nordbeck 1967, Nilsson 1968, Butler 1969). The latter two examples study how a part, or the whole, of a built-up area functions, seen from the user's point of view. Thus building function analysis has been extended to include town function analysis.

In the same way that building function analysis can be naturally extended to include town function analysis, so town function analysis can be extended to include region function analysis. The following example illustrates this. The construction of a shopping centre in the vicinity of a town affects the range of service facilities the town can provide. Shops in housing areas may become uneconomic and be closed. Thus the quality of function of the housing area is reduced seen from the user's point of view (Malmgren 1971). The quality of service provided for the individual house declines, except in the case where it lies in the immediate vicinity of the new shopping centre. A location in the immediate vicinity of the shopping centre can cause a decline in the quality of the housing area for other reasons, such as traffic noise resulting from the increased traffic the centre creates. The establishment of a shopping centre also affects the range of service provided in other towns. Thus one cannot limit studies to the town, the region as a whole must be studied. At this level it becomes increasingly necessary not to forget the individual user and the requirements as to function he has regarding his surroundings.

One can proceed a further stage up the hierarchy and study the relationships between regions, creating a subject which one could describe as

society function analysis. Even in this case the user lies at the heart of things and the subject looks at those aspects of how the society functions which interest the individual user.

One can also refer to levels in a hierarchy in conjunction with planning. Pure building function analysis is concerned with the micro-local level, town function analysis deals with the local level, region function analysis deals with the regional level while society function analysis deals with the national level in the hierarchy. The boundaries between the different levels are difficult to define but the terms local level, regional level and national level will be used in this report.

One should also be aware of the fact that it is impossible and also undesirable to separate planning at one level from planning at another. A decision taken to refer to the national level will, in most cases, also affect planning at the regional and local levels and vice versa.

If one accepts the extension of building function analysis to include parts of town, region and society function analysis then it is also natural that the use of maps becomes widespread even in this subject. Maps such as those which show the relationships between a user and his surroundings or the distribution of users within an area. The increased use of maps has created an interest in automatic map construction methods and this interest has resulted in a clearly expressed wish to take part in the development of automated mapping systems of the NORMAP type. It has also shown itself possible to apply programs from the NORMAP system when analysing relationships inside a building (Cole 1970) or in the immediate vicinity of a building (Dahlgren, Hellberg & Lindberg 1971). Hence NORMAP has a wider field of application than was thought from the beginning.

6 Location problems and NORMAP

The programs included in NORMAP can be classified into different sub-groups, for mapping, for approximation, for choice of location etc.

The most important mapping programs are those which construct isarithmic maps. In addition to mapping these programs can be used in the definition of boundaries such as the boundaries of built-up areas (Nordbeck 1969).

The location programs included in the system are of only, slightly less importance than the mapping programs and an example of their application has been included in part II of this report. The example is preceded by a description of the central place theory of amongst others Christaller (1933) and Lösch (1954) as applied to Swedish data. This is followed by

a presentation of a mathematical location model forming a part of the NORMAP system which can be used to determine the optimum locations and umlands for a number of service facilities within an area. This model pays greater regard to the user and to the way the user uses service facilities than the central place theory does.

The stability of the solutions given by the location model is studied with regard to inaccuracies in the input data and simplifications in the model. How accurately must the location and distribution of the customers be described to the model? how inaccurate can the input data be? how accurately should the user's costs incurred when visiting the facility be described? how accurately should the distance between the user and the facility be given? how are the results affected when regard is given to the redistributions of population which have occurred in Sweden since the middle of the 19th century? These are all questions which are studied in this context. The model is made more complicated later to allow the location of multiple facilities in stages, with those facilities located in earlier stages retaining a fixed location. The model is also capable of handling choice of locations for different groups of facilities simultaneously, i.e. simultaneous hierarchical location. A facility can be located independently or in combination with facilities from some of the other groups. Finally the model is used to define regions, i.e. the optimal location and umland is determined for a large number of facilities. These regions are then combined to form larger regions, every region consisting of a suitable number of sub-regions. The results obtained in this way are compared with those obtained by the opposite procedure, first the definition of the regions at the highest level in the hierarchy, then a decomposition into a number of smaller regions.

As one can see the location models are widely applicable. This has resulted in the wide use of different programs in the NORMAP system to prepare data for and to test the location models. Thus, the NORMAP system has had a very thorough testing.

Part I

The mapping system NORMAP

1 Computerized map construction

The suggestion that computers could be used in the making of maps was made by cartographers fairly soon after computers became generally available. Several cartographers began studying the application of computers to the laying out of maps, including the location and lettering of place names and the positioning and drawing of map symbols. If a topographical map is available in a numerical form then it can be manipulated by a computer; this further allows the combination of material from a variety of sources and the use of the computer to draw the finished map. This drawing is carried out on an automatic drawing machine (plotter) which can be directly coupled to the computer (on line), or fed using a paper or magnetic tape originating from the computer (off line). If the drawing machine is equipped for engraving then the map may be transferred directly to the printing plate. For coloured maps the computer may be used to carry out the colour separation and three separate plates prepared using the drawing machine.

Another area of mapmaking, where computers began to be used at the end of the 1950's, is in the actual construction of the map. The computer constructed the map on the basis of known values associated with given points or areas. The resulting map was printed on a teletype, or on a line printer if the latter was available. The main problem with this technique was to develop an output method which produced an easily readable map.

A third way in which a computer can be used is to let it collect unprocessed information from one or more of the official registers which are available in computer readable form, for example the population register or the real estate register. This information can then be manipulated and combined to produce a map tailored to the requirements of the problem at hand. The whole procedure from the extraction of information from a data bank to the presentation of the finished map has thus been automated. The technique used for the presentation of the map forms an important part of this process, but even more important are the methods used for processing the original information and for the actual construction of the map. The production of an easily read map is not sufficient if the methods

used to construct it are not accurate. Accuracy is of course relative, but in this case it is taken to mean that when tolerances are stated in advance, these tolerances should then not be exceeded in the actual construction of the map.

1.1 SYMAP

The best known of the different mapping systems is probably SYMAP (SYnographic MAPping system), developed originally at Northwestern University and perfected at Harvard (Fisher 1970). The presentation techniques it contains are highly developed, with the drawback that the only output device used is the line printer. On the other hand the map construction techniques it contains are less satisfactory. Output from SYMAP is built up from shaded areas, the precise type of shading used is open to decision by the user. SYMAP's shading technique is of very high quality and it is unlikely that it can be improved to any degree. This means that maps produced using SYMAP are aesthetically very attractive.

If a line printer equipped with differently coloured carbons is available then it can be used to produce coloured maps using the SYMAP system.

The shaded areas are built up by allowing up to four alphabetic characters to be printed over the top of one another. A white space is obtained if the "space" symbol is used. On the other hand if the alphabetic characters O, X, A and V are used the space becomes practically black.

The SYMAP system produces three main types of map, CONTOUR, CONFORMANT, and PROXIMAL. The first of these, CONTOUR, is an isarithmic map which should be used primarily for the presentation of such continuous variables as topography, rainfall or population density (Fisher 1970). The other main type of map, CONFORMANT, is built up of shaded areas where the size, shape and position of the different areas are given in the form of closed polygons. All qualitative maps are members of this extremely large group. In maps of this type all the points inside an area are assumed to have exactly the same value.

The SYMAP system has been installed at several Scandinavian universities (Bergström 1969—1972). It has even been used to check visually building coordinate points obtained from the Swedish real estate register (Bergström 1971b).

1.2 LINMAP

The English mapmaking system LINMAP (LINE printer MAPping) differs in many ways from the American SYMAP, but has the major similarity

that a line printer is used for output of the completed map. LINMAP was developed to handle information obtained from the UK census for 1966. This means that input data corresponding to districts or parishes can be presented in map form. The presentation techniques used appear less highly developed than those of SYMAP and the finished map presents a less satisfying visual appearance. The map construction methods used are also simpler than those of SYMAP. On the other hand the data processing techniques included in LINMAP are of a similar standard to those of SYMAP.

LINMAP can produce four different types of map, Dotmap, Gridmap, Zonemap and Surfmap (Gaits 1969). Dotmap differs from the usual type of dot map in that for every given point the associated value is written out in full. Gridmap is an ordinary square net map. Zonemap is similar to the CONFORMANT map produced by SYMAP, similarly Surfmap is equivalent to the SYMAP CONTOUR i.e. an isarithmic map.

Coloured maps can be produced using a version of LINMAP called COLMAP. This version produces a paper or magnetic tape suitable for an electronic typesetting machine. The resulting coloured map is produced by printing on a conventional printing press.

1.3 NORMAP

The Swedish mapmaking system NORMAP uses two output devices, line printer and automatic drawing machine. The system is designed to manipulate coordinate based information where the coordinates are highly accurate and information refers mainly to values at points. This means that if the available information refers to areas then a point must be chosen to represent the area (Nordbeck 1962). Information relating to areas can always be transferred to points in this way.

At the moment work is in progress on a coordinate based real estate register for the whole of Sweden. This register will probably be completed during the first half of the 1970's. The coordinates for every real estate unit are given by choosing at least one point and giving it a coordinate value based on the coordinate system used in the Swedish land use map (1938 version). This coordinate system has the equator as one axis and a line 1500 km west of the meridian $2^{\circ}15'$ west of Stockholm's old observatory as the other axis.

The coordinate based real estate register and the population register can, by use of electronic data processing, be combined to form a data bank. The NORMAP mapping system is designed to accept information from

this databank and use it to construct several different types of map. Thus it can be said that both the information processing and map construction components of NORMAP are highly developed.

1.4 NORMAP's programming languages

The programs forming the NORMAP system are available in most cases in both ALGOL and FORTRAN versions. Some of the ALGOL programs have already been published (Bengtsson & Nordbeck 1964, Nordbeck & Rystedt 1967, 1969a, 1969b), and it would therefore appear unnecessary to deal with them here. The remaining ALGOL programs are written in UNIVAC NU-ALGOL and are therefore not directly convertible for use on other computers. As the majority of computers are now equipped with a FORTRAN compiler no particular limitations to the general use of NORMAP are created if publication is limited to the FORTRAN version (FORTRAN IV).

2 Terms, concepts and notation

A consistent cartographic terminology does not exist at the moment. As a result the terminology used in computer cartography varies widely, with large numbers of terms having only small differences in meaning (Nordbeck & Rystedt 1970). In order to avoid misunderstandings the most important terms used in conjunction with NORMAP together with their definitions are collected in this chapter. These definitions are straightforward and the terms used throughout this book correspond with these definitions. This has not been done in order to force a particular terminology on the reader, but in order to help the reader to follow the succeeding chapters.

2.1 Function

The term function is used in its mathematical meaning, i.e. it refers to an expression of the type Y is a function of X or more concisely $Y=f(X)$. As in most cases a map refers to a function which is defined over a plane the expression is of the type $Z=f(X, Y)$ where X and Y are the coordinates for a point P which lies on the function's definition plane. Examples of functions of this type are height above sea level, air pressure, temperature and rainfall per year.

2.2 Program, procedure and subroutine

The majority of the programs included in NORMAP exist in both FORTRAN and ALGOL versions. They are in most cases programmed as subprograms which can easily be included in a larger and more complicated main program. This type of subprogram is called a procedure in ALGOL and a subroutine in FORTRAN. In this publication these two terms are used synonymously, although the term procedure is used more often than the term subroutine. As both a procedure and a subroutine are a form of computer program the term program is used to describe both a

procedure and a subroutine, although of course the term program has a much wider meaning.

2.3 Constant, parameter and variable

The terms constant, parameter and variable are defined in accordance with information processing practice. A variable is therefore a quantity which changes as the calculation proceeds. The expression $Z=f(X, Y)$ contains three variables, X and Y are coordinates and Z is the value of the function at point (X, Y) .

A parameter is a quantity which is given as input to the computer and which is constant throughout that particular execution of the program; the program does not modify the value assigned to a parameter. The user can, of course, assign a new value to the parameter the next time the program is executed. A constant differs from a parameter in that it assumes the same value for every execution of the program.

The following simple example illustrates the difference between a variable, a constant and a parameter. If the diagonal of a square is represented by D and the area of the square by A the value of A can be determined from the expression $A=0.5 D^2$. In this expression A and D are variables and 0.5 is a constant. The expression can easily be extended to apply to all regular polygons and circles, in this case the expression becomes $A=CD^2$. If C is assigned the value 0.5 one obtains the area of a square, if the value $3\sqrt{3}/8$ (≈ 0.64) the area of a hexagon. If C is assigned the value $\pi/4$ (≈ 0.79) the area of a circle with a diameter D is obtained. As C assumes different values for different executions of the program it is a typical parameter.

The arguments used when a subroutine or procedure is called are often called parameters. A parameter of this type does not always obey the rule that its value is unchanged throughout the execution of a program, since calculated values can be assigned to a parameter of this type which will cause its value to change.

2.4 Coordinates

The coordinates used by Swedish planners are normally those of the coordinate system used by the Swedish land use map. This system has one axis parallel with the equator and the other parallel with a meridian 2.5° (2.5 new degrees) west of Stockholm's old observatory. In order to ensure

positive coordinates throughout Sweden the origin point of the system is placed on a line parallel with and 1500 km west of the above mentioned meridian. As a result of this every point in Sweden can be described by a coordinate containing the same number of figures, independent of whether it is a north—south or east—west coordinate, with the reservation that both the coordinates are of the same degree of accuracy.

The number of figures in a coordinate is dependent on the accuracy chosen, if the coordinates are given to the nearest kilometre then a 4 figure coordinate is obtained, if to the nearest metre then a 7 figure coordinate is the result.

It should be observed that the east—west coordinates always begin with 1 inside Sweden whereas the north—south coordinates begin with 6 or 7. In accordance with mathematical practice the east—west coordinate is represented by X and the north—south by Y . This is a reversal of the arrangement used by surveyors. Property coordinates are given according to surveying practice and need therefore to be reversed before manipulation with NORMAP. This change can be carried out automatically when the coordinates are read by the computer. In the expressions and calculations presented in this report X always refers to the east—west coordinate and Y to the north—south.

2.5 Grid point and grid net

A grid point is a point for which the value of the function to be mapped is known or for which it is to be calculated. Grid points are always nodes in a grid net. The meshes' shapes in a net of this type can be completely irregular but they are more often rectangular or triangular. The most common form of net is triangular. The triangles used in such a net can be of all three types, irregular, isosceles or regular. If one regards an ordinary road network as a grid net then the road junctions are the grid points and the blocks are the meshes. The actual streets correspond to the lines which join the grid points. The grid net most often used with NORMAP is triangular with the base of the triangle equal to its height. However, several of the NORMAP programs are designed to accept both triangular and rectangular grid nets.

A grid point is represented by the letter G plus a subscript, e.g. G_i . In those situations where one is concerned within which row or column a grid point lies a double subscript is used, e.g. G_{rc} , in this case r refers to the row and c to the column. Occasionally one requires the coordinates for a grid point, in this case the notation (X, Y) is used together with a subscript i.e. (X_i, Y_j) or (X_c, Y_r) .

2.6 Data point

In most cases of map construction a function is used for which the values are known at several grid points arranged in a grid net. The actual net can be either regular or irregular. In the latter case a regular grid net is often placed over the irregular grid net and the value the function takes at the new grid points calculated using the values known for the points in the irregular net. When used in this way the grid points in the irregular net function mainly as data bearers. This function explains the alternative term for points of this type, data point. A grid point and the functional value associated with it are used for the actual construction of the map, a data point and its associated data value are used when determining the functional value associated with a grid point.

2.7 Reference interval and reference area

It is seldom enough to use one and only one point to determine the value of the function at that point. This is illustrated by the following example: 15 mm of rain fall between 1 o'clock and 7 o'clock. The time interval is necessary if the quantity of rain is to have any meaning. In this case the time between 1 o'clock and 7 o'clock is a reference interval and the example is one of a typical reference interval function.

A statement of the type "his income is 2000 kr", is completely meaningless without its corresponding reference interval. On the other hand the statement "his monthly income is 2000 kr", is well defined and shows that income is another example of a reference interval function. The majority of functions presented in map form are reference interval functions, for example population density, which is the number of persons who live in a given reference area. In those situations where the reference area is of varying size the associated values require adjustment before they can be compared with one another. This is done in a straightforward fashion by dividing the value for each area by the size of the reference area expressed in a basic unit of area. It should be observed that this unit of area should not be very much larger than the smallest reference area as otherwise very misleading results may be obtained (Nordbeck & Rystedt 1970).

2.8 Equidistance

In those situations where a regular triangular or rectangular grid net is used the grid size is given by the distance between two successive grid

points in the same row. This distance is called the equidistance for the grid net and is usually represented by the letter H . In the triangular grid net often used with NORMAP the triangles are isosceles with the base of the triangle equal to the height, both being equal to H . In this net every alternate row of grid points is displaced half the equidistance relative to the adjoining rows.

2.9 Overlapping

The only way of raising the accuracy of a map is by the use of a finer grid net in its construction. Increasing the number of grid points can lead to an overlapping of the reference areas. If a circle of diameter D is used as the reference area then an overlapping parameter S can be defined as in expression 2.1. If D is used to refer to the length of side of a

$$S=D/H \quad (2.1)$$

reference square then the number of grid points associated with a reference square is S^2 .

3 Square net maps and the square grid net

The simplest type of map which can be constructed using a computer is the square net map. This type of map can be produced as output using a line printer and is basically a table. Hannerberg (1961) emphasises this characteristic of the square net map by calling it a chorological matrix. Unfortunately the output from a line printer or teletype is not easily organized into a square arrangement. The vertical equidistance of a table produced on a line printer is often 6 steps per inch and the horizontal equidistance often 10 steps per inch. Converted to millimetres these dimensions are 4.23 mm and 2.54 mm respectively. The smallest square one can produce is thus 3 steps in the vertical direction and 5 in the horizontal direction. If a multiple of the basic square is not applicable then some form of approximation must be used.

This is in contrast to the situation when an automatic drawing machine, a plotter, is used as output device. As any size of rectangle can be drawn using this type of device scale problems are avoided, since the step size of such a plotter can be as small as 0.1 mm. However, because of speed differences, it may be necessary to output the whole of the map on a line printer in order to select especially interesting sections for drawing on the plotter.

3.1 The square net map program NORK

This program uses a square grid net and is a fast program which should be used as widely as possible.

3.1.1 Variables used in the NORK program

(X, Y) are the coordinates for a data point P . $DATA$ is the data associated with point P . R is the row number of a cell in the map. C is the column number of a cell in the map.

The row number is counted from the bottom edge of the map upwards beginning with 1. The column number is counted from the left hand edge of the map to the right beginning with 1.

3.1.2 Parameters used in the NORK program

(X_0 , Y_0) are the coordinates of the map's origin point, i.e. its lower left hand corner.

L is the length of the map expressed in rows.

W is the width of the map expressed in columns.

H is the equidistance. As the squares do not overlap one another H is the same as the size of the cells.

MAP in the ALGOL programs is a 2-dimensional array containing the actual map. A cell in the map, with row number R and column number C is identified by the notation $MAP(R, C)$. In the FORTRAN programs MAP is a 1-dimensional array. The cell $MAP(R, C)$ above is in this case identified by $MAP((C-1) \cdot L + R)$.

LUN is the logical unit number, usually of a binary tape from which NORK obtains the data to be processed to produce the map.

3.1.3 The mathematical principles used in NORK

The computer begins by reading in a record containing the coordinates for one point and the data associated with it. The information is then assigned to its respective cell in the map. This is carried out by calculating R using formula 3.1 and C using formula 3.2. The operation INT means that only the integer part of the result of the bracketed expression is relevant.

$$R = INT((Y - Y_0)/H) + 1 \quad (3.1)$$

$$C = INT((X - X_0)/H) + 1 \quad (3.2)$$

A check is then carried out to see that the cell chosen lies inside the area to be mapped i.e. that R is larger than or equal to 1 and smaller than or equal to L and that C is larger than or equal to 1 and smaller than or equal to W . The final stage is the adding of the information to the actual map. See formula 3.3.

$$MAP(R, C) = MAP(R, C) + DATA \quad (3.3)$$

The next record is then read in and handled in the same way and so on until the available information is exhausted. The map can then be outputted using program MAPOUT, stating in advance that the output is required in square net map form.

Example: The Fortran statement `CALL NORK (1345, 6245, 20, 32, 5, MAP, 4)` causes a square net map to be constructed. The south-west corner of the map has the coordinates (1345, 6245) and the map contains 20 rows and 32 columns. The cell size is 5 km by 5 km and the map shows therefore an area 100 km long and 160 km in the east—west direction. The last parameter indicates that the value attached to *LUN* is 4. This logical unit which usually refers to a magnetic tape must be allocated data in some way. This can be carried out by using subroutine *MAPIN* as described in paragraph 3.4. Another method is the use of a special excerpt program *NORXRP* to extract information from a central databank, which is then transferred to magnetic tape.

3.2 The overlapping program NORKO

3.2.1 The parameters used in the NORKO program

The NORKO program uses the same type of square grid net as the NORK program. The starting point for these two programs is thus similar and they in fact use the same parameters with the addition of a further two in the case of NORKO, i.e. *RFORM* and *S*.

NORKO uses an overlapping technique where a reference area overlaps other reference areas. The overlapping constant *S* indicates the extent of overlapping, if *S* is equal to 1 then no overlapping occurs.

The length of side used for the reference square is equal to $H \cdot S$. The parameter *RFORM* is used to describe the shape of the reference area. If *RFORM* is given the value 0 then the reference area is taken to be circular with a diameter of $H \cdot S$ units. With another value for *RFORM* e.g. 4 the reference area is taken to be a square.

3.2.2 The influence area method

NORKO works in a similar way to NORK in that the *X* and *Y* coordinates of a data point plus its associated data *DATA* are read in. Those grid points which lie inside an influence area surrounding this data point are then determined and the value *DATA* added to the values associated with each of them.

An influence area is nearly identical with a reference area as is shown by figure 3.1. A square reference area does not contain its upper edge or its right hand side, this reference area is obtained automatically when the techniques contained in NORK are used to obtain the row number *R*

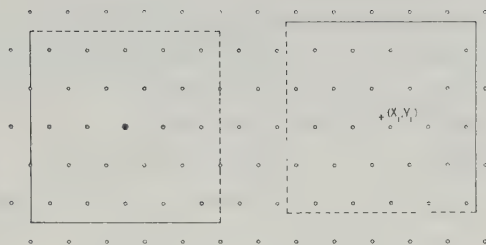


Figure 3.1. A reference area placed over a grid point (left hand diagram) and the corresponding influence area placed over point (x_i, y_i) .

and the column number C . The corresponding influence area is obtained by reflecting the reference area on its north-west south-east diagonal. This means that an influence area does not contain its left hand and bottom edges. The influence area method gives the same result as the more traditional manual approach, when a reference area is placed over a grid point, and the data points which lie inside the reference area are determined and their value added to that of the grid point. The manual method can be used as the basis for a computer program but the result is very slow and ineffective.

3.3 The MAPOUT program

This program is used to output a numerical map generated by NORK or NORKO on a line printer. To carry out this operation values are required for the following parameters, $X0$, $Y0$ (the coordinates of the origin of the map), L (the length of the map), W (the width of the map), H (the equidistance), and MAP . In addition a value is required for the parameter $GFORM$ which is used to describe the type of grid net used. If $GFORM$ is given the value 3 the output is given in the form of a triangular grid net with the same equidistance between grid points in a row as between the rows themselves i.e. H . Every other row is displaced half this equidistance making every triangle isosceles with the same height and length of base.

When outputting a triangular grid net the map's equidistance must be a multiple of 1 inch. In the case of a square grid the equidistance is required to be a multiple of 0.5 inch. In the version of MAPOUT used at present triangular output is produced with an equidistance equal to 0.5 inch. In this case the displacement of alternate rows is not completely accurate, 3 steps instead of 2.5 but the result is visually fully acceptable. Every value printed takes up 5 positions, even though in certain cases several of these positions are blank. Between every line of values printed there are two blank lines. If the equidistance is required to be other than 0.5 inch then

the number of positions occupied by every value must be increased. If for example every value is allocated 10 positions then the equidistance becomes 1 inch. In this case the number of blank lines between every line of values must be increased from 2 to 5.

MAPOUT also calculates the length of 1 coordinate unit of the outputted map and prints it on every new sheet together with the values of the other parameters. In addition the X and Y coordinates of the rows and columns are given along the edges of the map.

3.4 The MAPIN program

MAPIN is used to input information in the form of a square net map for further manipulation with for example the NORKO program. It thus constructs a data file where every record consists of an X and an Y coordinate together with a value $f(X,Y)$. This is done by entering the values for more than one cell on each punched card. The first three values on a card are the coordinates (X,Y) for the first cell in a row together with its value $f(X,Y)$. The following values are $f(X+1,Y)$, $f(X+2,Y)$, $f(X+3,Y)$ etc., that is the value in the second, third, fourth etc. cell in the row. When a punched card is full entries are continued on the next card beginning with the coordinates for the next cell to be entered.

The only parameter used when calling MAPIN is LUN . This parameter refers to the logical unit via which the completed data file is to be outputted. The other parameters are read in from punched cards during the execution of MAPIN.

4 Regular triangular grid nets and isarithmic maps

The numerical maps produced by NORKO are in many cases all that is required, but occasionally a more easily read map is needed which gives a better presentation of the total situation. In these cases an isarithmic map is often used. The word isarithm is built up from two greek roots, *isa* meaning equal and *ritm* meaning value. Another name for isarithm could therefore be equal-value-line. Preparation of an isarithmic map can be carried out manually, using as the starting point a numerical map of the type produced by NORKO. When this manual procedure is carried out a number of rules of thumb are applied. These rules are difficult to describe in terms of a computer program when transferring the manual method to a computer. A method for constructing isarithmic maps with the computer's help differs therefore from the corresponding manual approach, in that it is based on the computer's mode of operation. This provides an example of the fact that computer cartography is not just the transfer of existing manual methods to a computer program.

4.1 Maps and 3-dimensional diagrams

Every map can be treated as a 2-dimensional reproduction of a 3-dimensional diagram. Thus for every type of diagram there exist a corresponding map. In the case of a 3-dimensional histogram the diagram is built up of a number of pillars, the area of the bottom of the pillar is the same as the corresponding reference area, the volume of the pillar is proportional to the value to be shown for the area in question. The height of a pillar is therefore proportional to the value of the function at that point, i.e. the value to be shown divided by the size of the reference area. Construction and presentation of this type of histogram requires a great deal of effort and it is therefore often replaced by the use of a shading technique. Each reference area is then given a tone lying on a grey scale proportional to the value to be shown. Usually this is organized so that the higher the

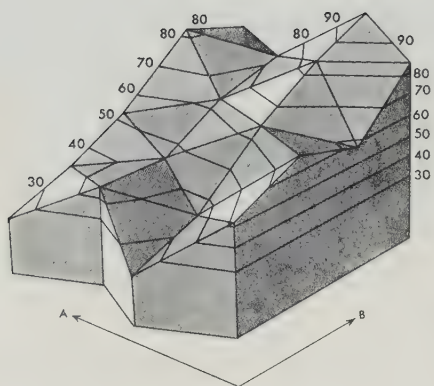


Figure 4.1. A 3-dimensional diagram, the upper surface of which has been intersected with planes parallel with the base plane.

value associated with a reference area the darker the tone. The SYMAP and LINMAP mapping systems use this shading method widely.

4.2 Polygon diagram and isarithmic maps

A 3-dimensional polygon diagram corresponds to an isarithmic map in the same way that a shaded map corresponds to a 3-dimensional histogram. In a 2-dimensional polygon diagram the curve showing how the function varies is a polygon train. In the 3-dimensional polygon diagram the line sections are replaced by triangles. The surface of this diagram is thus built up from triangles as is shown in figure 4.1.

The 3-dimensional polygon diagram has two major disadvantages; first that it is complicated, needing much time and effort for its construction and second that it is difficult to determine from the diagram the value of the function at a particular point. In figure 4.1 a method for increasing the readability of this type of diagram is shown. Horizontal section lines are added at the junction of the upper surface of the diagram with equally spaced planes lying parallel with the diagram's base. These lines are actually lines of equal value which, when projected on to the plane at the base of the diagram, give an isarithmic map (figure 4.2).

4.3 Location of the isarithms by interpolation

The method for location of isarithms described in the previous section is of course far too complicated for practical use. As can be seen in figure 4.1 an isarithm lying on the surface of a triangle is a segment of a straight line.

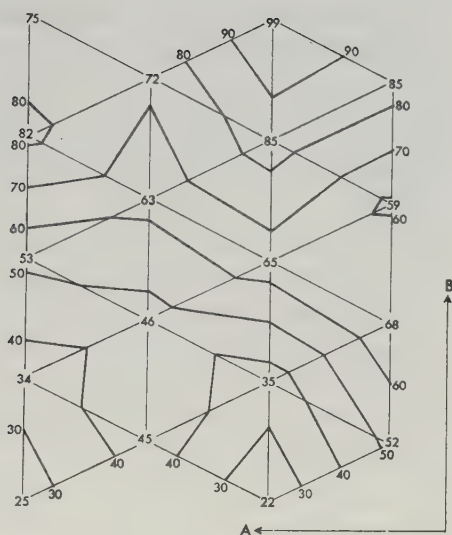


Figure 4.2. The lines of intersection between the planes and the upper surface of the diagram shown in figure 4.1, projected on to the base plane.

Thus if the end points of the isarithm segment are known the position of the whole segment is known. As a result it is only necessary to interpolate on the lines forming the sides of the triangle when determining the position of the isarithms.

The coordinates for two grid points which completely determine one side of a triangle can be given the notation (XU, YU) and (XV, YV) respectively. In the same way, the function value at these two points can be given the notation ZU and ZV respectively. A further point (XI, YI) lies on the side of the triangle and on an isarithm of value ZI . The quotient D is determined using expression 4.1 and the values of XI and YI determined using expression 4.2.

$$D = (ZI - ZV) / (ZU - ZV) \quad (4.1)$$

$$\begin{cases} XI = XV + D(XU - XV) \\ YI = YV + D(YU - YV) \end{cases} \quad (4.2)$$

The above expressions assume that the value ZI lies between value ZU and ZV . The next stage involves the use of the third point in the triangle in order to determine the position of the other end of the isarithm. The procedure is repeated for every one of the isarithms required which go through the triangle and for every one of the triangles which form the grid net. The result of these operations is that every isarithm is located in the form of unconnected line segments.

4.4 The joining up of line segments to form isarithms

If an isarithmic map is all that is required the line segments can be drawn out to produce a satisfactory result. On the other hand if the isarithms are required in the form of linked polygon trains further operations are necessary.

Once one has located the beginning of an isarithm following it is a straightforward matter, especially when a triangular grid net is used. Beginning at the junction of an isarithm with the side of a triangle the next point on the isarithm must lie on one of the other sides of the triangle but this second side also forms the first side of another triangle and the procedure can therefore be repeated. In this way the line segments can be built up into linked polygon trains.

4.5 The NORI program

When constructing isarithmic maps, the grid net ought to be built up of regular triangles, but this requirement clashes with the facts that the co-ordinates of the grid points ought to be integers and that it ought to be possible to cover the whole area of the map with reference squares without overlapping occurring. This third requirement has as its result the same equidistance between columns as between rows.

A grid net which satisfies the second and third requirements and which is built up of triangles as regular as possible has every other row of grid points displaced half the equidistance H (Nordbeck 1964b). The resulting triangles are isosceles with the height of the triangle equal to the base, i.e. equal to H . The other two sides are as a result $\sqrt{5}/2 H \approx 1.1 H$.

The above grid net should be used when preparing a numeric map for further processing with the isarithmic interpolation program NORIP. The NORI program uses a net of this type but is in other features similar to NORKO. However when printing a map constructed with the help of NORI on the line printer the parameter *GFORM* in the *MAPOUT* program must be allocated the value 3.

4.6 The interpolation program NORIP

4.6.1 The parameters and variables used in the NORIP program

The following parameters are included in the procedure call for NORIP; *X0*, *Y0*, *L*, *W*, *H*, *MAP*, *GFORM*, *LUN* and *NETH* which are all

integers and *COUNT* and *EKVID* which are real parameters. The first 7 of these parameters locate a grid net and the function values at the grid points. The values stored in *MAP* should be equal to or larger than 0 and smaller than 100 000. If the numerical map stored in *MAP* has been constructed using a program based on a triangular grid net, for example *NORI* the *GFORM* should be given the value 3. On the other hand if the grid net stored in *MAP* is based on a square grid *GFORM* should be given the value 4. As a triangular grid net is always used in constructing an isarithmic map *NORIP* modifies the rectangular net to give a triangular one by the addition of suitable diagonals as shown in figure 4.3.

If the parameter *LUN* has a value greater than 0 it describes the logical unit to which the result is to be transferred. This logical unit cannot refer to either the line printer or the punched card punch. The results given via *LUN* can be used as input data to the plotting program *NORIPL* which draws the completed isarithmic map on the plotter. The parameters *NETH*, *COUNT* and *EKVID* can be given arbitrary values if *LUN* is greater than 0.

If *LUN* is equal to 0 the isarithms are not stored but plotted out directly. In this case *NETH* is the equidistance of a square net plotted out at the same time to simplify the reading of the completed map. The parameter *COUNT* is assigned the length in centimetres of 1 *H* on the ground and the parameter *EKVID* is given the length in centimetres which 1 *H* is to assume on the map. The scale of the map is thus the quotient of *EKVID/COUNT*.

When *NORIP* is called during the execution of a program the parameters *K* and *ISAC* are read from punched cards. *K* is the number of isarithm levels to be calculated, and *ISAC* is an array containing the different values for which isarithms are to be calculated i.e. *ZI* in expression 4.1.

In order to be able to use the same interpolation method when an isarithm passes directly over a grid point the value of the function at this point is increased by a very small value *EPS* which is greater than 0. This results in the isarithm passing very slightly to one side of the grid point. The resulting inaccuracy is too small to be detected by the eye in the finished map.

R and *C* refer as usual to the row number and column number but they can also be treated as coordinates of the grid point in question. In the same way *A* and *B* can refer to another grid point which together with point (*R,C*) describes the line along which interpolation is being carried out. When carrying out the next interpolation (*A,B*) becomes the first point (*R,C*) while a new point is chosen to take the place of (*A,B*). The choice

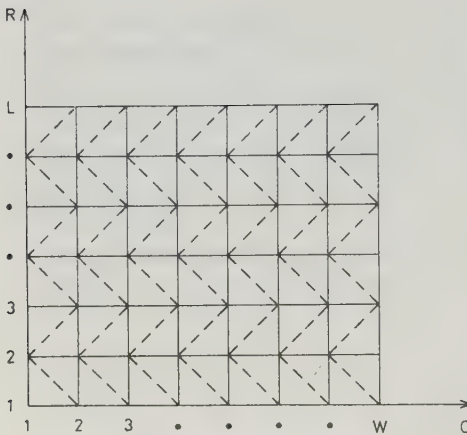


Figure 4.3. A rectangular grid net can be converted to a triangular grid net by the addition of diagonals in the manner shown.

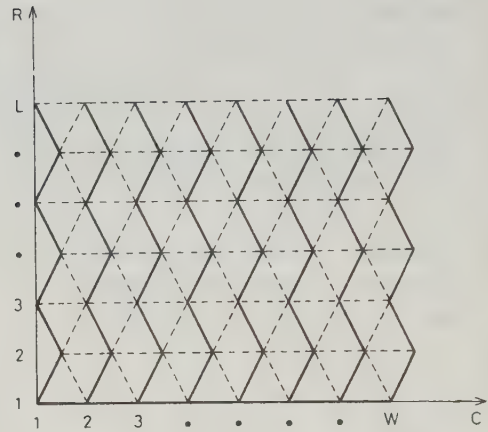


Figure 4.4. Interpolation along the solid lines in the diagram guarantees that all isarithms are treated.

of this new point is carried out with the help of two logical variables *EVEN* and *DELTA*. When interpolation is being carried out along an even row *EVEN* assumes the value *true*, when working with a triangle lying upright *DELTA* also becomes *true*.

4.6.2 The interpolation method used in NORIP

A check is carried out to see if the first isarithmic value $ZI = ISAC(1)$ is less than $MAP(1,1)$ which is the value of the function at the first grid point in the first row of the map. If this is so the edge of the map is followed until a grid point with a lower value than that of ZI is discovered. The point where the isarithm crosses the edge of the map is determined by linear interpolation. From this point the isarithm can be followed from triangle to triangle to the point where once again it crosses the edge of the map. This edge is included in the isarithmic polygon train. This inclusion is carried out by checking grid points along the edge of the map, beginning at the intersection of the isarithm and the edge, for values higher than ZI . When a value lower than ZI is discovered it means that a new isarithm has been crossed. The isarithm is then followed into the map. This procedure is repeated until one has returned to the lower left hand corner of the map and the isarithm is described in the form of a closed polygon.

If, on the other hand, ZI is larger than $MAP(1,1)$ the search continues by checking if ZI is smaller than $MAP(1,2)$. If it is, then an isarithm must pass

between them. This isarithm can then be followed in the way described above. The program proceeds along the lower edge of the map, checking each pair of grid points in this way and following the isarithms as they occur.

When the lower edge of the map has been completed in the manner described the search continues along the left hand edge of the map and for all the vertical zig-zag lines as shown in figure 4.4. The result of the above is that every isarithm is crossed at least once. In order to avoid treating the same isarithm more than once the grid points on each side of an isarithm are marked by adding a large figure (100 000) to the value of the function at each grid point. This means that every grid point value must be larger than or equal to 0 and smaller than 100 000 and this is tested at the same time as the comparison with the value of *ZI*.

4.7 The transformation routine MAPTRF

The NORIP program cannot be used to construct an isarithmic map if any of the functional values are less than 0 or larger than 99 999. MAPTRF transforms maps containing these values so that they satisfy the requirements of the NORIP program. This transformation is carried out by adding to all the values the smallest power of 10 which makes the lowest negative value positive. If as a result of this addition certain of the values have risen over 99 999 the program divides all the values by that power of 10 which reduces the highest value to under 99 999. The resulting map can then be processed further using NORIP.

The call for procedure MAPTRF contains three parameters, *L* the number of rows in the map, *W* the number of columns in the map and *MAP*.

4.8 The storage of numerical maps on magnetic tape

It is often necessary to store a numerical map on magnetic tape. This map can originate from a number of programs including NORK, NORKO and NORI. The subroutine which carries out this operation is called MAPTAP. The subroutine call contains the following parameters, *X0*, *Y0* (the co-ordinates of the origin of the map), *L* and *W* (the length and width of the map), *H* (the equidistance used for the grid net), *MAP*, *GFORM* which describes the shape of the grid net and *LUN* which is the logical unit number on which the array *MAP* is to be stored.

5 Irregular grid nets

Data is often recorded associated with points distributed in the form of an irregular grid net. In many cases it is also desired to present this information in the form of an isarithmic map. An example of this type of data is population data at the parish level which is often considered to be associated with the point at which the church is situated. Lines can be drawn between these points (churches) to give an irregular grid net. If this net is to be used to construct an isarithmic map its meshes must be triangular.

The net containing weather recording stations as its nodes is a further example of this type of irregular grid net. Such functions as temperature, rainfall, and air pressure are often shown with help of an isarithmic map which has been constructed on the basis of an irregular triangular grid net.

Churches and weather recording stations are fairly stable institutions and the same grid net can therefore be used a large number of times. Thus sufficient motive exists for the development of a computerized technique for the preparation of isarithmic maps from an irregular triangular grid net.

An example of another type of irregular nets is that used when measuring a stretch of country in order to construct a contour map. It is highly unlikely that the same net would be used when carrying out a later survey. Due to the fact that it is a difficult and slow job to describe a grid net to a computer a method has been developed for producing isarithmic maps from points in an irregular grid without the necessity for describing the actual grid. The two approaches to this problem are the transformation of the irregular grid to a regular grid and the use of the points in the irregular grid as data points from which the value of the function at points in a new, regular grid is calculated. Transformation of the irregular grid requires the description of the grid in some way, which then makes possible the application of NORIPO, the interpolation program for irregular grids. This approach will therefore not be handled here.

5.1 Approximate calculation of functional values at grid points in the regular grid

The value of the function is known at a number of points spread irregularly over an area. In order to separate these points from the points in the regular grid they can be called data points. On the basis of these data points the value of the function at the grid points in the regular net can be calculated. These values can then be used as input to the NORIP program. The calculation of the approximate values can be carried out using two main groups of methods, interpolation methods and approximation methods.

5.1.1 Interpolation in LINMAP and SYMAP

Examples of different interpolation methods can be found in the two mapping systems described earlier, LINMAP and SYMAP. In the LINMAP system independent data points lying separate from the regular grid cannot be used. The grid points are organized in a regular net based on the dimensions of the line printer output. The problem is thus to calculate values for every grid point for which data is not available which is done by linear interpolation along both rows and columns. The mean value of these two approximations is then used as function value in the grid point.

SYMAP uses a considerably more advanced interpolation method than LINMAP. When calculating the value of the function at a grid point the 4—10 data points which lie nearest the grid point are used. The value of the function at everyone of these data points is then multiplied by a weight which is inversely proportional to the distance from the grid point to the data point. These products are then summed to give the value of the function at the grid point. In many cases the distance used in this calculation is not the actual physical distance but a power of it. This interpolation technique is very similar to the calculation of population potentials. The further away a data point is from the grid point the less effect it has on the value of the function at the grid point.

LINMAP and SYMAP do not construct the isarithms in the form of polygon trains. Instead the isarithms are presented as blank lines between areas of different tones in the final print out. It is therefore necessary to make sure that the isarithms do not lie too close to one another i.e. a shaded area must lie between two isarithms and the blank lines together must form a continuous isarithm.

5.1.2 The NORINT program for interpolation

A grid point G has the coordinates (R,C) . It lies in a mesh of an irregular grid which is defined by the points $D1$, $D2$, and $D3$. These points are data points and have the coordinates $(X1,Y1)$, $(X2,Y2)$, and $(X3,Y3)$. The value of the function at these points is given in a similar way by $Z1$, $Z2$, and $Z3$, while the value of the function at the grid point is given the notation Z . The interpolation is carried out by passing a plane through the three points $(X1,Y1,Z1)$, $(X2,Y2,Z2)$, and $(X3,Y3,Z3)$ and by determining its equation. The coordinates of the point (R,C) are then entered into the equation and the value of Z determined. Due to the fact that the 3-dimensional interpolation formula is exceedingly long it is not given here. The interested reader is referred to program NORINT where the above formula is used. Unfortunately it is difficult to determine which data points should be used as the basis for the interpolation. If the irregular net has been fully described then a point-in-polygon program, for example NORPCONVEX (Nordbeck & Rystedt 1967) can be used or alternatively program NEARLINK (Nordbeck & Rystedt 1969b). In most cases one desires to avoid needing to describe the irregular net in a detailed form, since this is a very large job and the data points themselves are more important than the links which join them. However, one can always use the three data points which lie nearest grid point G , these points thus become $D1$, $D2$, and $D3$ in the interpolation formula. The data points which form the basis for interpolation are defined by program NORID. Use of this method avoids the problems which occur when the irregular net has to be described fully.

5.1.3 Approximate determination of functional values with program NORIAP

The NORINT program has one serious drawback, that only three data points are used in the interpolation. The NORIAP program on the other hand uses a different principle. It fits a function of the type $Z = A + BX + CY + DX^2 + EY^2$ to a series of points, $(X1,Y1,Z1)$, $(X2,Y2,Z2)$, $(X3,Y3,Z3)$, $(X4,Y4,Z4)$, . . . using the least-square method. It is obvious that the number of points used should not be too few, there should be at least as many points as there are parameters in the equation. In this case the number of parameters is 5, i.e. A , B , C , D , and E . Therefore at least 5 points should be used. The data points to be used for approximation are, as is the case with NORINT, determined by application of the NORID program.

There is of course no reason why a function of a type other than $Z = A + BX + CY + DX^2 + EY^2$ cannot be used when carrying out this approxi-

mation. A function of the type $Z = A + BX + CY$ could also be applied using in this case three data points, although a faster result would be obtained by using the NORINT program.

WEIGHT is the only real parameter which is used in the NORIAP program. It states whether the value of a data point is to be weighted in proportion to the distance between the grid point and the data point. If this parameter is given the value 0 no weighting is carried out, otherwise the expression $DIST^{-WEIGHT}$ where *DIST* refers to the distance between the two points is used as the weight.

5.1.4 Preparation of the input data for the NORINT and NORIAP programs

These two programs both require access to a fast method for determining which data points are to be used to approximate the value of a function at a grid point. The best approach would appear to be the use of the NORI program to build up a list of data points in the order of their distance from a grid point and to carry this out for every point in the grid. In practice the list contains the *X* and *Y* coordinates for the data point, the value of the function at that point and the distance from the data point to the grid point in question. This distance is stored in such a way that the data point which lies nearest a grid point is placed first in the list for that particular grid point, the next nearest second and so on. If the result from the application of NORI is that there are three points in this list for every grid point then these can be used directly in NORINT. If more than three are present then obviously only the first three are used. If on the other hand less than three are present the NORI program is called again using an increased diameter for the reference circle. On the second, or subsequent, iterations those points which are already included in each list are not treated.

The parameters *MIN* and *MAX* indicate the minimum and maximum number of data points which are to be included in each list. In order to conserve storage space *MAX* cannot be larger than 10. If NORI finds more than 10 data points within the reference area those lying furthest from the grid point are disregarded.

The method described above forms the basis of the NORID program. This program must therefore be called before applying either NORINT or NORIAP. The resulting data points calculated by NORID are stored on a magnetic tape which then forms the input data for NORINT and NORIAP.

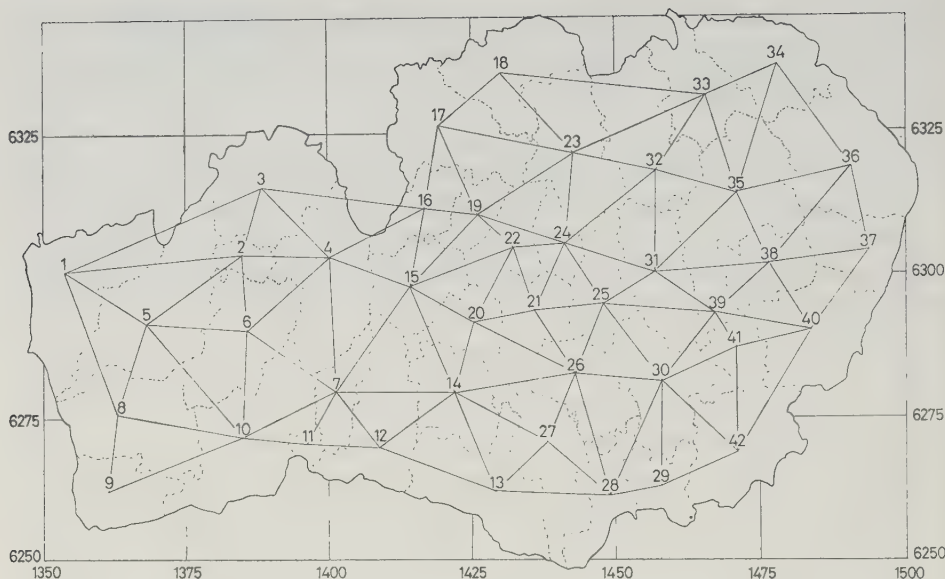


Figure 5.1. An irregular triangular grid net.

5.2 Storage of irregular grid nets

An irregular grid net is stored using a method developed originally for the study of road networks (Nordbeck 1964a, 1969). A grid net can be understood as a series of nodes (street corners) and links. The first stage is the numbering of the data points, grid points in the irregular net, using consecutive numbers, see figure 5.1. A list is then prepared giving the number of each data point, I , its coordinates ($X(I), Y(I)$), together with the number of each of the points which are direct neighbours. The number of direct neighbours is not allowed to be larger than 7. Table 5.1 describes completely the irregular grid net shown in figure 5.1. It should be noticed that every link occurs twice in the table, once as link IJ and once as link JI .

5.3 The NORIPO program for irregular grid nets

The NORIPO program constructs an isarithmic map using an irregular grid net. The procedure begins by determining which links in the irregular net are crossed by an isarithm of a given value. This is done by checking whether the value of the isarithm lies between the functional value at the

No	X	Y	Immediate neighbours						
1	13540	63001	8	5	2	3			
2	13850	63032	1	3	4	6	5		
3	13884	63150	1	2	4	16			
4	14002	63026	2	3	6	7	15	16	
5	13683	62912	1	2	6	8	10		
6	13860	62898	2	4	5	7	10		
7	14012	62792	4	6	10	15	11	12	14
8	13630	62752	1	5	10	9			
9	13614	62618	8	10					
10	13850	62713	5	6	7	8	9	11	
11	13962	62700	7	10	12				
12	14090	62696	11	7	14	13			
13	14291	62620	12	14	27	28			
14	14221	62790	7	12	13	27	20	26	
15	14144	62976	4	7	14	20	16	19	22
16	14169	63112	3	4	15	19	17		
17	14192	63256	16	18	23	19			
18	14300	63348	23	17	33				
19	14260	63101	15	16	17	22	24	23	
20	14254	62914	14	15	22	21	26		
21	14361	62936	20	22	24	25	26		
22	14324	63044	15	19	20	21	24		
23	14427	63208	17	18	19	24	32	33	
24	14410	63050	19	21	22	25	31	32	23
25	14478	62946	21	24	26	31	30	39	
26	14429	62835	14	20	21	25	27	30	28
27	14381	62707	13	14	26	28			
28	14489	62612	13	27	26	29	30		
29	14576	62626	28	30	42				
30	14578	62810	25	26	28	29	42	41	39
31	14566	63000	24	25	32	35	38	39	
32	14568	63178	23	24	31	35	33		
33	14654	63306	18	23	32	35	34		
34	14776	63360	33	35	36				
35	14708	63136	31	32	33	34	36	38	
36	14902	63180	34	35	37	38			
37	14932	63038	36	38	40				
38	14762	63014	31	35	36	37	39	40	
39	14670	62930	30	25	31	38	40	41	
40	14836	62900	37	38	39	41	42		
41	14708	62868	30	39	40	42			
42	14710	62687	29	30	41	40			

Table 5.1. The irregular grid net shown in figure 5.1 given in table form, with the number of every grid point, its position, (X and Y coordinates) and the numbers of its immediate neighbours.

grid point and the functional value at its neighbour. In order to avoid identifying each link twice only those neighbours with a number higher than that of the grid point are used.

Every link crossed by the isarithm is recorded by its end points in two 1-dimensional fields, $NB1$ and $NB2$. When every grid point has been

treated every stored link refers to one junction between a link and an isarithm. It now remains to determine the location of every junction and to combine the junction points into a polygon train.

The location of every junction point is determined by linear interpolation using expressions 4.1 and 4.2. The construction of the isarithmic polygon begins with the first link in the field. This link forms a side in two different triangles and of these two one is chosen and interpolation continued using the method shown in figure 5.2. In this way two successive vertices of the isarithmic polygon have been located. This procedure is repeated for the succeeding triangles until the starting point is reached. The isarithm is thus defined in the form of a closed polygon. If the isarithm crosses the outer polygon in the grid net it must do so in two places. The computer includes in the polygon that section of the outer polygon which lies between these two crossing points, see figure 5.2. This procedure makes all the isarithms into closed polygons which simplifies their output on the plotter and their use when calculating areas and volumes.

Immediately after every link has been used in the construction of a polygon its storage positions in fields *NB1* and *NB2* are given the value 0. An exception to this rule is made in the case of the first link which is necessary to determine the position of the last point in the isarithmic polygon. Often there is more than one isarithm with the same value in the area covered by the map. This eventuality is checked by seeing if the two fields *NB1* and *NB2* are empty when an isarithm has been constructed. If positions remain to be used then further isarithms at the same level must exist. This further isarithm is constructed in the same manner as the first and the procedure repeated until no further links remain. The operation is then repeated for the isarithms at the next level.

EPS is a small number, which is added to the value of the function at a grid point in the event that this value is exactly equal to the value of the isarithm under construction. This technique allows the same interpolation method to be used in all situations.

One of the few restrictions which apply to the program NORIPO is that the number of vertices in the isarithm polygons under consideration should not be larger than 5000. This number is to a certain extent dependent on the number of links which is itself dependent on the number of nodes *N* in the grid net. The number of links is usually approximately $3N$ with at the most every alternate link crossed by an isarithm. Thus NORIPO can deal with grid nets containing up to 3500 grid points and in most cases even more.

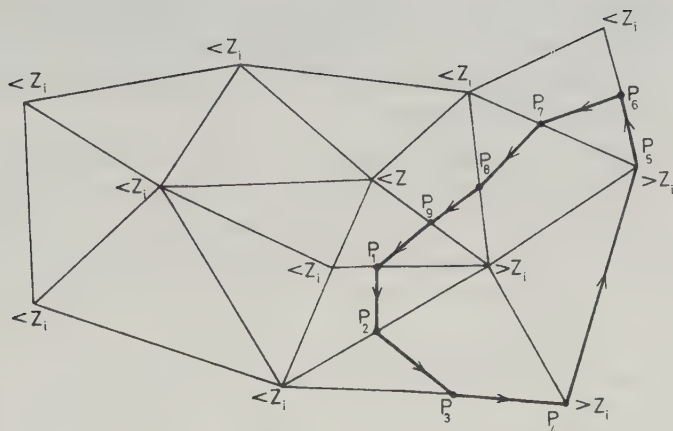


Figure 5.2. The first point on isarithm Z_i , which is encountered by the NORIPO program, is referred to as P_1 . The program then follows the isarithm and describes it in the form of a closed polygon. Grid points with a higher value than Z_i lie immediately to the left of the isarithm.

5.4 Transformation of a regular grid net into an irregular grid net

A comparison between the two isarithm programs NORIP and NORIPO shows that NORIP is approximately 10 times faster than NORIPO. NORIP on the other hand has a larger number of restrictions. The value of the function cannot be larger than 99 999 or less than 0 for example. It can therefore in some cases be advisable to use NORIPO in conjunction with a regular grid net. In this case the regular net is handled precisely as if it were an irregular net. A regular grid net can of course be described in the same way as an irregular net, i.e. the number of every grid point, the coordinates for each point, the value of the function at each point and the numbers of a grid point's immediate neighbours.

This transformation can best be performed by the computer and a procedure has therefore been written to carry out this operation with the name NORTRO. The output from NORTRO can be used directly as input to NORIPO.

5.5 The graphical output procedure NORIPL

NORIPL draws polygon trains and closed polygons on a line plotter. Every vertex in a polygon train is described by three variables, (X, Y, Z) .

X and Y are the coordinates of the point in question and Z is a number by which the polygon train is identified. This program is particularly suitable for drawing isarithmic maps and in this case Z is usually given the value of the isarithm in question.

The call on NORIPL contains 8 parameters, the coordinates of the lower left hand corner ($X0, Y0$), and those of the top right hand corner ($MAXX, MAXY$) come first, followed by $NETH$ which is the equidistance of a grid net drawn by the plotter. The procedure does not check to see if any part of the isarithm polygons lies outside the area of the map as defined by the first 4 parameters.

The real parameter $COUNT$ gives the dimension of the basic coordinate unit in centimetres on the ground, $EKVID$ gives the dimension of this increment in centimetres on the map. Hence, the scale of the map is the quotient $EKVID/COUNT$.

$EKVID$ is restricted by the requirement that $(MAXY - Y0) \cdot EKVID$ should not be larger than the maximum width of the paper in the plotter. In the X direction no corresponding restriction exists.

The LUN parameter refers to the logical unit used for storing the isarithms forming the map, usually the same logical unit is used as was used when storing the completed polygons from NORIP and NORIPO.

The plotter begins by drawing a rectangle to contain the map and by drawing the net with graduations along its sides. The first polygon is then read in via LUN and drawn out. All the isarithms for each level are stored together and all the isarithms at one level are therefore drawn in succession. The plotter then proceeds to draw out the next level and so on.

6 The calculation of volumes and areas

6.1 Polygons and the determinant formula

A polygon with n sides has n corners, $P_1, P_2, \dots, P_n$. The corner point P_i has the coordinates (X_i, Y_i) and all polygons are described in counter-clockwise order, i.e. they are positively orientated, see figure 6.1.

If the coordinates for the vertices of a triangle are known the area can easily be calculated using the determinant formula 6.1.

$$2T_3 = \begin{vmatrix} X_1 & X_2 & X_3 & X_1 \\ Y_1 & Y_2 & Y_3 & Y_1 \end{vmatrix} = X_1Y_2 + X_2Y_3 + X_3Y_1 - (Y_1X_2 + Y_2X_3 + Y_3X_1) \quad (6.1)$$

The above formula can be stated in an alternative manner containing three simple determinants (formula 6.2).

$$\begin{aligned} 2T_3 &= X_1Y_2 - Y_1X_2 + X_2Y_3 - Y_2X_3 + X_3Y_1 - Y_3X_1 \\ &= \begin{vmatrix} X_1 & X_2 \\ Y_1 & Y_2 \end{vmatrix} + \begin{vmatrix} X_2 & X_3 \\ Y_2 & Y_3 \end{vmatrix} + \begin{vmatrix} X_3 & X_1 \\ Y_3 & Y_1 \end{vmatrix} \\ &= |P_1P_2| + |P_2P_3| + |P_3P_1| \\ &= \{P_1P_2P_3P_1\} \end{aligned} \quad (6.2)$$

A triangle is of course a polygon and one can therefore ask the question, is there a determinant formula which applies to all polygons? to which the answer is yes. Formula 6.3 shows this general version which can be proved by using the inductive method. One accepts that formula 6.3

$$2T_n = \{P_1P_2P_3 \dots P_nP_1\} \quad (6.3)$$

is true for a polygon with n sides and shows that it also applies to an $(n+1)$ -sided polygon. It is already known that it is true for a triangle which can be proved by calculating the area of a triangle using three trapeziums in a right angled coordinate system. Hence, the determinant formula 6.3 also applies to tetragons and to pentagons, and so on.

The proof is based upon the division of an $(n+1)$ -sided polygon, $P_1P_2 \dots P_{n+1}P_1$, into an n -sided polygon $P_1P_2 \dots P_nP_1$, and a triangle

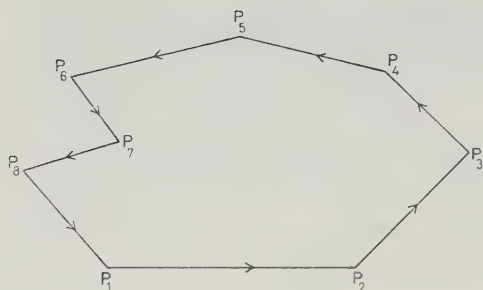


Figure 6.1. A positively orientated polygon.

$P_1P_nP_{n+1}P_1$, as shown in figure 6.2. The area of an $(n+1)$ -sided polygon can be referred to as T_{n+1} which can be determined from the area of a triangle T_3 and the area of an n -sided polygon T_n . The above discussion can be stated as in expression 6.4.

$$2T_{n+1} = 2T_n + 2T_3 \quad (6.4)$$

In formula 6.5 the areas T_n and T_3 have been replaced by the equivalent determinant notation used in expression 6.3.

$$2T_{n+1} = \{P_1P_2 \dots P_nP_1\} + \{P_1P_nP_{n+1}P_1\} \quad (6.5)$$

The right hand side of the expression is then divided up into simple determinants as in 6.2 to give formula 6.6.

$$2T_{n+1} = |P_1P_2| + |P_2P_3| + \dots + |P_{n-1}P_n| + |P_nP_1| + |P_1P_n| + |P_nP_{n+1}| + |P_{n+1}P_1| \quad (6.6)$$

One can see directly from this expression that the determinants $|P_1P_n|$ and $|P_nP_1|$ contain the same columns. The well known rule, that only the sign of a determinant is changed if two rows or columns in it are shifted thus applies. These two determinants thus cancel out one another to give the following formula 6.7

$$2T_{n+1} = |P_1P_2| + \dots + |P_{n-1}P_n| + |P_nP_{n+1}| + |P_{n+1}P_1| \quad (6.7)$$

The right hand side of formula 6.7 can be transcribed using the rules given earlier to give a result for an $(n+1)$ -sided polygon equivalent to expression 6.3. The formula is therefore proved to apply to all polygons.

A criticism of the above method is that the division into an n -sided polygon and a triangle was not carried out randomly. The triangle can be considered special therefore in that it is built up of the points P_1 , P_n , and P_{n+1} . On the other hand the numbering of the points is completely random and any one of the points can be given the notation P_1 . One can commence by dividing the $(n+1)$ -sided polygon into an n -sided polygon and a triangle and then

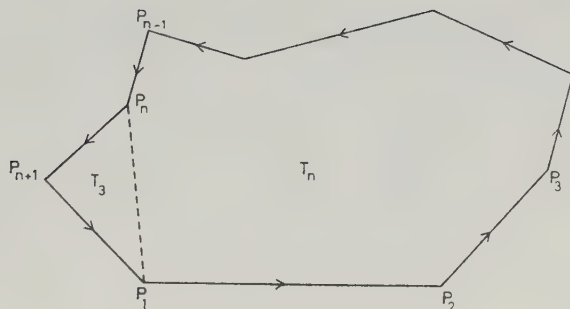


Figure 6.2. An $(n+1)$ -sided polygon can always be divided up into an n -sided polygon (T_n) and a triangle (T_3).

carry out the numbering of the points. This criticism is not therefore justified.

If one traverses the periphery of the polygon in the reverse direction, i.e. clockwise, and applies the above formula the resulting area is negative. A further result obtained when using formula 6.3 is thus the orientation of the polygon. This fact can be used when determining whether a point lies inside or outside a polygon and makes formula 6.3 one of the most important components in the point-in-polygon programs NORP and NORPCO.

6.2 The calculation of volumes using Simpson's formula

NORYT is a fast procedure used to calculate the area of n -sided polygons. This fact makes it possible to calculate the area of every isarithm polygon at the same time as the isarithm is developed without noticeable increase in the computer time required. NORYT is in this case used in conjunction with NORIP or NORIPO. An isarithmic map is a 2-dimensional representation of a 3-dimensional diagram. Any 3-dimensional diagram has a volume which is proportional to the volume of the object mapped. A 3-dimensional diagram showing a piece of landscape, a hill for example, can thus be used to calculate the actual volume of the hill in reality. A more interesting fact is that an isarithmic map of the hill can also be used to calculate the volume. A requirement for this is that the area contained by each isarithm is known. Both NORIP and NORIPO give their isarithms as closed polygons, the orientations of which are determined by the rule that the value of the isarithm is always lower than the value of the function at points lying immediately to the left of the isarithm. As a result of this rule when an isarithm describes a hill it is positively orientated, when describing a depression it is therefore negatively orientated. The total area

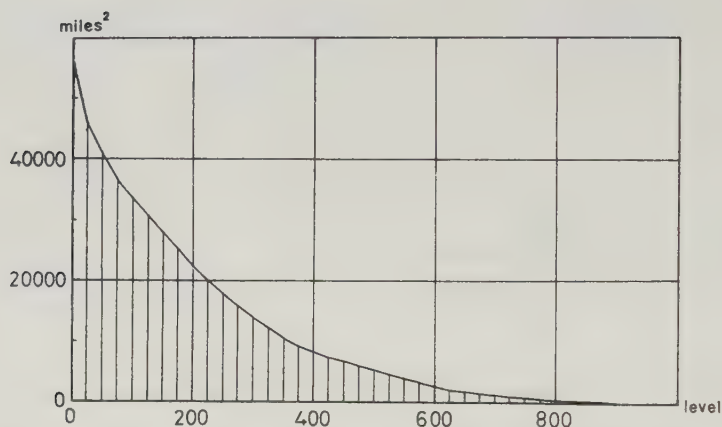


Figure 6.3. A diagram showing the total area of the isarithm polygons at different isarithm levels. The levels are given in thousands of persons per 2500 square miles. The volume of the isarithmic map, i.e. the total population of the area mapped is directly proportional to the shaded area under the curve and can be simply calculated on the basis of it.

bounded by an isarithm at a particular level is obtained by adding together the area of all the polygons at that level paying attention to the sign of each polygon. If the area of each polygon has been determined through use of the determinant formula the correct sign is obtained automatically.

The distance between two adjoining isarithm levels is equidistant and is therefore referred to as HN . The area at isarithm level number i is referred to as A_i . These two values can be arranged as a point in a graph having the isarithm level as its x -axis and the values of A_i as its y -axis, see figure 6.3. If the equidistance HN is small these points in the graph will tend to form a continuous curve. The area between this curve and the x -axis is thus equal to the volume of the isarithmic map.

It is unfortunately not possible to use a value of HN which is equal to 0. The curve must therefore be approximated and the simplest way of carrying out this approximation is by joining every point to the next with a straight line, thus forming a polygon train. The area under this curve can be calculated by the use of formula 6.8. V refers to the volume of the map which is the same as the area bounded by the two axes and the curve itself.

$$V = \frac{HN}{2} (A_1 + 2A_2 + 2A_3 + \dots + 2A_{n-1} + A_n) \quad (6.8)$$

Formula 6.8 is based on repeated calculation of the area of a trapezium. Simpson's formula is a more complicated but also more accurate method

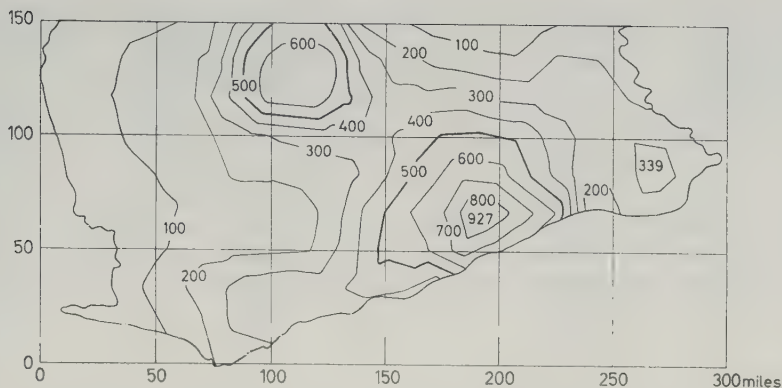


Figure 6.4. An isarithmic map showing the distribution of population in southern Ghana in 1960. Data: Claeson, Godlund & Örtenblad 1969.

for carrying out the same calculation, see formula 6.9. When using this formula the number of isarithm levels should be odd. If this is not

$$V = \frac{HN}{3} (A_1 + 4A_2 + 2A_3 + \dots + 4A_{n-1} + A_n) \quad (6.9)$$

the case a further level of area $A_{n+1} = -A_n$ is used. In most cases the area of A_n is small thus allowing $A_{n+1} = 0$ to be used without causing noticeable inaccuracy. Correction with 0 is easiest to use from a computer programming viewpoint.

6.3 Volume calculations — a numerical example

The discussion above concerning the volume of isarithmic maps applies to all types of isarithmic map, not only to topographical maps. Figure 6.4 shows an isarithmic map of population distribution in southern Ghana in 1960, the data for this map is obtained from a square net map of the area (Claeson, Godlund & Örtenblad 1969).

The reference area is a square with sides 50 miles long. The isarithmic map shows isarithms for 100 000, 200 000, . . . , 800 000 inhabitants per reference square. Table 6.1 shows the area bounded by each isarithm in increments of 25 000 persons per reference square. The 0 isarithm can be approximated by using the 1000 isarithm as the difference between these two areas is exceedingly small.

The areas calculated by NORYT are given in units determined by the user and stated by the value given to parameter *COUNT*. In the Ghana

Isarithm level in thousands per 2500 square miles	Area in square miles	Isarithm level in thousands per 2500 square miles	Area in square miles
1	56 777	475	6 199
25	45 866	500	5 490
50	40 755	525	4 789
75	36 667	550	4 108
100	33 604	575	3 473
125	30 782	600	2 791
150	28 067	625	2 338
175	25 729	650	1 926
200	22 607	675	1 554
225	20 203	700	1 220
250	17 808	725	948
275	15 804	750	797
300	14 085	775	659
325	12 486	800	532
350	10 793	825	419
375	9 530	850	314
400	8 510	875	216
425	7 662	900	126
450	6 914	925	44

Table 6.1. Areas of isarithmic polygons given in square miles. The different isarithm levels refer to number of persons, measured in thousands of persons per 2500 square miles. The reference area is a square with each side 50 miles long. This example refers to Ghana. See also figure 6.3.

example the equidistance H is 2 which in reality corresponds to 16 $\frac{2}{3}$ miles. In order to obtain the areas calculated in miles² *COUNIT* requires the value 8.333 . . . , in this case the overlapping parameter S is 3. The actual reference area R is thus a square with sides 50 miles long with an area of 2500 miles². The isarithm equidistance HN is given in inhabitants per reference square, in this example 100 000 inhabitants/ R .

The volume of the isarithmic map, i.e. the total population is obtained by using the areas in table 6.1 together with the trapezium formula 6.8 or Simpson's formula 6.9.

With HN equal to 100 000 inhabitants per reference area use of the trapezium formula gives a total population of 4 694 000 persons in southern Ghana. Simpson's formula used on the same data gives a population of 4 583 000 persons. The actual population of the area in question is 4 529 000, showing that Simpson's formula gives a slightly more accurate result, see table 6.2.

The difference between the correct population and that calculated from the volume of the corresponding isarithmic map is relatively small. A better approximation would have been obtained if a smaller equidistance

Equidistance	Volume Simpson	Volume Trapezium
200 000	4 705 000	5 026 000
100 000	4 583 000	4 694 000
50 000	4 538 000	4 577 000
25 000	4 530 000	4 542 000

Table 6.2. The volume of the isarithmic map shown in figure 6.5. The volume has been calculated in two ways, with Simpson's formula and with the trapezium formula. The volumes should be compared with the actual population in southern Ghana which is 4 529 000.

between the isarithms had been used. If HN is changed to 50 000 the trapezium formula gives a population of 4 577 000, which can be compared with the 4 538 000 persons given by Simpson's formula. A further reduction of HN to 25 000 persons gives a further improved result, 4 542 000 for the trapezium formula and 4 530 000 for Simpson's formula.

The volume of an isarithmic map can be fairly easily arrived at, with only two calculations, for HN and for $HN/2$. This is possible due to the fact that when using Simpson's formula the total inaccuracy when using HN is approximately 16 times as large as that when using $HN/2$. In the Ghana example the inaccuracy was 9 000 persons with HN equal to 50 000 persons and 600 with HN equal to 25 000 persons giving a quotient for the inaccuracy of approximately 15. These figures cannot be completely trusted as inaccuracies caused by rounding off are also involved.

6.4 The NORVSI program for volume calculations

The NORVSI procedure calculates the volume of an isarithmic map using the simpler trapezium formula and the more complicated Simpson's formula. In both cases correction with 0 is used i.e. the isarithm at level $N + 1$ is given as having a zero area. The array A contains the area bounded by the isarithms at each level together with the value of the isarithm. N refers to the number of isarithm levels and HN to the equidistance between them. The results are allocated to parameter VT in the case of the trapezium formula and VS in the case of Simpson's formula.

6.5 Numerical maps and the calculation of volumes

If the numerical map which forms the basis of a particular isarithmic map is available it can be used directly to calculate the volume. This is particularly simple when using a grid net built up from isosceles triangles as is

the case with NORI. A numerical map originating from NORI can, as an alternative to being used in the construction of an isarithmic map, be used to construct a 3-dimensional diagram of the type shown in figure 6.5. This diagram consists of a number of triangular pillars, the three parallel sides of which are referred to as Z_1 , Z_2 and Z_3 . The base of every pillar is formed of an isosceles triangle of area $H^2/2$ where H is as usual the grid net's equidistance. The volume of each pillar can be calculated using expression 6.10.

$$VP = H^2(Z_1 + Z_2 + Z_3)/6 \quad (6.10)$$

Every Z value applies to 6 different pillars, thus the volume of this type of numerical map can be calculated by summing all the Z values and multiplying by the square of the equidistance as in equation 6.11.

$$V = H^2 \sum_{r=1}^L \sum_{c=1}^W Z_{rc} \quad (6.11)$$

A formula corresponding to 6.11 can also be developed for irregular grid nets. It is more complicated than formula 6.11 and it is therefore more economical to transform the irregular net into a regular net by use of the NORINT or NORIAP programs. These programs interpolate the Z values at the grid points in a regular grid thus allowing formula 6.11 to be used in calculating the volume which is done by aid of the procedure NORV.

6.6 Numerical terrain models

An interesting application of isarithmic techniques and the calculation of volumes is provided by numerical terrain models. A survey of a stretch of landscape provides information in the form of numerical triples (X, Y, Z) where Z is the value of the function at point (X, Y) . These triples create a numerical model of the terrain surveyed which can be manipulated in several ways. A common approach is to use them to calculate the Z values at the grid points in a regular grid net. This transformed model can then be used when drawing sections through the terrain. One can describe isarithms as horizontal profiles constructed with the help of existing vertical profiles. The construction of an isarithmic map is thus considerably more difficult than the construction of the corresponding vertical sections.

Formula 6.11 as embodied in the NORV procedure can be used when calculating the volume lying between 2 distorted surfaces. There is in

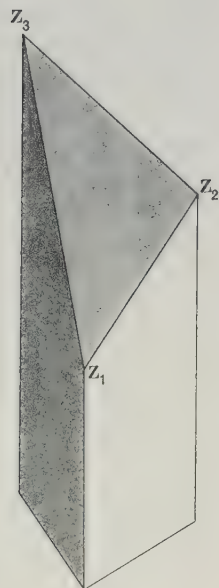


Figure 6.5. A 3-dimensional diagram of the type shown in figure 4.1 actually consists of a number of pillars. The volume of pillars of this type can be simply calculated using the formula $B(Z_1 + Z_2 + Z_3)/3$ where B refers to the area of the base.

principle no requirement that the lower of these two surfaces should be a plane or some form of water level, i.e. Z can be defined in the case of two distorted surfaces as the distance between the two. A practical way of using this technique is when a numerical terrain model is used to represent a planned landscape arrangement. The volume lying between the planned surface and the existing surface can then be calculated. This volume represents the amount of excavated material to be transported to or from the site. As this operation is expensive it is desirable to reduce its quantity to a minimum. A further item of information which is often required is the amount of excavated material to be moved inside the actual area, this can be obtained using the positive Z values independently. It is highly likely that the first planned surface is unsatisfactory and that changes to it are carried out in order to reduce the amount of excavated material moved.

Certain modifications are required to the model of the existing terrain, in order to allow for the material to be excavated from the actual foundations of the building. This takes the form of an additional mound of material of the same volume as the planned foundation which is placed at the correct location in the model.

A numerical terrain model can be combined with a mathematical model, in that the planned surface is replaced by a mathematically defined sur-

face. A sloping plane defined by an equation of type $Z = A + BX + CY$ could be used. In this case the choice of values for A , B and C which minimize the movement of the excavated material is a simple mathematical problem. The use of surfaces with more complicated equations is also possible. Use of this approach when planning an area can lead to considerable savings which more than compensate for the computer costs involved.

7 The fitting of a polynomial to a set of points

7.1 The fitting of a plane to a set of points in a 3-dimensional space

In the case of a numerical terrain model the value of Z is known at a number of points (X,Y) . Z can refer to a height over sea level, in this case the triple (X,Y,Z) refers to a point in a 3-dimensional space. A number of these points N build together a set of points, a point swarm. The problem is to fit a surface to this point swarm which lies as near to all the points as possible. The simplest and best known way of carrying this out is by the least-square method.

$$Z = A + BX + CY \quad (7.1)$$

The general form for the equation of a plane is given by equation 7.1. The values of the coefficients A , B , and C are to be chosen so as to bring the plane as near as possible to all the points (X_i, Y_i, Z_i) . When the least-square method is used, the coefficients A , B and C are determined in such a way that the function $F(A,B,C)$ in equation 7.2 is minimized.

$$F(A,B,C) = \sum_{i=1}^N (A + BX_i + CY_i - Z_i)^2 \quad (7.2)$$

Partial derivation of F with respect to A , B and C with the partial derivatives equal to 0 gives equation system 7.3. The values of A , B and C are then determined by solving this equation system.

$$\begin{aligned} A \sum_{i=1}^N 1 + B \sum_{i=1}^N X_i + C \sum_{i=1}^N Y_i &= \sum_{i=1}^N Z_i \\ A \sum_{i=1}^N X_i + B \sum_{i=1}^N X_i^2 + C \sum_{i=1}^N X_i Y_i &= \sum_{i=1}^N X_i Z_i \\ A \sum_{i=1}^N Y_i + B \sum_{i=1}^N X_i Y_i + C \sum_{i=1}^N Y_i^2 &= \sum_{i=1}^N Y_i Z_i \end{aligned} \quad (7.3)$$

7.2 Truncation errors and the mean transformation

The coefficients of equation system 7.3 can easily become very large numbers. If X is a 6 figure number the square of X extends to 12 figures. The required addition of N of these 12 figure numbers can easily lead to very large numbers. If N is approximately 1000 the resulting coefficients can be up to 15 figures long. Even larger coefficients are obtained when fitting a second degree surface to the point swarm. In this case with N approximately 1000 and X 6 figures long the resulting coefficients can have up to 27 figures. Such numbers cannot be handled by a computer without truncation errors occurring. A computer incapable of handling values more than 18 figures long accepts the first 18 figures in a large number and gives the remaining figures the value 0. The resulting equation system can therefore become totally incorrect as the lowest figures of the coefficients are in fact the most important. In order to avoid these truncation distortions the whole coordinate system can be transformed to place its origin in the vicinity of the mean point (XM, YM, ZM) . This is done by replacing X by $X - XM$, Y by $Y - YM$ and finally Z by $Z - ZM$.

This type of translation leaves the values of B and C in equation system 7.3 unaffected. On the other hand A becomes 0 when XM , YM and ZM are used. The correct value of A for the original coordinate system can be obtained by substitution in equation 7.4. In this equation XM , YM and ZM

$$A = ZM - B \cdot XM - C \cdot YM \quad (7.4)$$

refer to the exact mean values of X , Y and Z . Unfortunately when carrying out the transformation described above these values are not suitable. This follows from the fact that if the X , Y and Z values are integer then the transformed values should also be integer. This requirement can be satisfied by rounding off each mean value to the nearest integer. These new values are referred to as MX , MY and MZ respectively. They are used for the transformation and subsequent calculations, thus avoiding possible rounding off errors. On the other hand these errors are less serious than truncation errors and in most cases they cancel one another out.

7.3 The coefficient table, the general case

A study of equation system 7.3 shows that the coefficients can be calculated using a straightforward scheme as shown in table 7.1. At the head of this table one finds the 4 values 1, X , Y , and Z . In the left hand column one finds just 1, X and Y . The contribution of every triple to each

	1	X	Y	Z
1	1	X	Y	Z
X	X	X ²	XY	XZ
Y	Y	YX	Y ²	YZ

Table 7.1. Scheme for the calculation of the coefficients in equation system 7.3, for the case with 3 variables, (X,Y,Z).

coefficient is then obtained by multiplying a value from the head of the table with the corresponding value in the left hand column.

The coefficient table 7.1 applies when a plane is being fitted to a collection of points. This method can on the other hand be made more general in its application. Assuming that instead of triples (X,Y,Z) one has a set of (T+1)-tuples (X₁,X₂,X₃,...,X_T,Y) where T refers to the number of independent variables. In this case the aim is to fit a function of the type shown by equation 7.5 as nearly as possible to all the points (X₁, X₂,X₃, ..., X_T,Y).

$$Y = A_0 + A_1 \cdot X_1 + A_2 \cdot X_2 + A_3 \cdot X_3 + \dots + A_T \cdot X_T \quad (7.5)$$

The method of doing this can be shown in a general table using the same principles as table 7.1. At the head of this table, one finds the T+2 values 1, X₁, X₂, X₃, ..., X_T, and Y; in the left hand column the same values occur again with the exception of Y. This table 7.2 can be used when calculating the coefficients of the equation system in a similar way to table 7.1.

This approach can also be used when the formula is not of the simple linear type as represented by equation 7.5. If X₃ is assumed to be equal to X₁², X₄ equal to X₂² and X₅ equal to X₁ · X₂ the resulting surface is of the second degree, having the equation shown at 7.6. The technical

$$Y = A_0 + A_1 \cdot X_1 + A_2 \cdot X_2 + A_3 \cdot X_1^2 + A_4 \cdot X_2^2 + A_5 \cdot X_1 \cdot X_2 \quad (7.6)$$

problem involved in fitting a second degree surface to a collection of points is not different to that occurring when fitting a plane. Restrictions do not occur even when fitting third-degree surfaces or when using surfaces having more than two dimensions such as hyperplanes and second degree hypersurfaces.

7.4 Logarithmic transformation

It is in many cases possible to convert a complicated equation into the general type as illustrated by equation 7.5. An example of the above

	1	X1	X2	X3	..	XT	Y
1	1	X1	X2	X3	..	XT	Y
X1	X1	X1 ²	X1X2	X1X3		X1XT	X1Y
X2	X2	X2X1	X2 ²	X2X3		X2XT	X2Y
X3	X3	X3X1	X3X2	X3 ²		X3XT	X3Y
..	..						...
XT	XT	XTX1	XTX2	XTX3		XT ²	XTY

Table 7.2. The general scheme for the calculation of coefficients for the case with $T + 1$ variables. The table is used when fitting a curve using the least-square method.

procedure is that practiced when using a gravity model, see equation 7.7. In this equation the interaction between two localities is referred to as I while P_k and P_l refer to the population of locality k and l respectively. D_{kl} is the distance between locality k and locality l .

$$I = A_0 \cdot P_k^{A_1} \cdot P_l^{A_2} \cdot D_{kl}^{A_3} \quad (7.7)$$

Equation 7.7 is often used in traffic generation models. In such case I describes the number of journeys to work from locality k to locality l . P_k would in this case refer to the size of the working population at locality k while P_l refers to the number of people employed at locality l .

Use of the gravity formula in the form given by equation 7.7 is extremely awkward and a transformation is therefore used. The transformation is carried out by taking the logarithm of both sides of equation 7.7 to give an equation of the general type, see formula 7.8.

$$\log I = \log A_0 + A_1 \log P_k + A_2 \log P_l + A_3 \log D_{kl} \quad (7.8)$$

Table 7.2 can then be used to construct the equation system which when solved gives values for constants A_1 , A_2 and A_3 directly. It is on the other hand necessary to take the antilogarithm of A_0 in order to obtain the actual value.

7.5 The solving of equation systems

Several methods exist for solving linear equation systems. One of these is the Gauss' elimination method (Fröberg 1962). This method was programmed for computer by Andersen (1962) using the procedure name LLGAUSS. Robertsson (1964) has improved LLGAUSS and it is this version with further improvements which forms part of the NORMAP system.

7.6 Correlation calculations

When the equation system coefficients have been arrived at, it is a simple matter to calculate the correlation coefficients for every pair of variables. In order to do this a 2-dimensional array is required for storing the correlation coefficients. This array is referred to as *RHO*. The correlation coefficients are calculated with help of equation 7.9, in this equation the expressions $CV(R,C)$, $CV(R,R)$ and $CV(C,C)$ are included which are calculated by using equation 7.10. $KO(R+1,C+1)$ is the element in row $R+1$ and column $C+1$ of the coefficient table 7.2.

$$RHO(R,C) = CV(R,C) / \sqrt{CV(R,R) \cdot CV(C,C)} \quad (7.9)$$

$$CV(R,C) = KO(R+1,C+1) - KO(R+1,1) \cdot KO(1,C+1) / N \quad (7.10)$$

7.7 The curve fitting program NORMKA

NORMKA fits a function of type $Y = A_0 + A_1X_1 + \dots + A_TX_T$ to a set of points given in the form of M -tuples $(X_1, X_2, \dots, X_T, Y)$. M is thus the number of variables ($M = T + 1$). N is the number of sets which should be larger than M .

NORMKA contains a procedure TRANSF with parameters *MV* and *LOG*. *MV* informs the program whether or not a mean transformation is to be carried out. If $MV = 0$ no transformation is carried out and the original values are used, when $MV = 1$ the procedure calculates the mean value of every variable and rounds it off, placing the resulting M values in an array *TR1*. Giving *MV* the value 2 causes M numbers to be read into the array *TR1*. The actual transformation is carried out by subtracting the respective value in *TR1* from each value of the corresponding variable. If it is desired to add to the value of a particular variable then the value read into *TR1* is given a negative sign. If *LOG* is given the value 1 a logarithmic transformation is carried out on some or all of the variables. The variables to be transformed are identified by values given to array *TR2*. A 0 in a particular position in this array indicates that the corresponding variable is not to be transformed, a 1 that the logarithmic transformation applies to that particular variable. Care should be taken that all the variables have positive values. If the original data contains negative values then a *MV* transformation must be carried out adding a suitable number to every one of the variable values. It must also be observed that the mean transformation is done before the logarithm transformation. Hence, the combination $MV = 1$, $LOG = 1$ is not permitted.

NORMKA works in stages beginning with function $Y = A0 + A1X1$. The accuracy of fit is measured by use of the residual variance RV and the residual standard deviation RS . The residual variance RV is used together with VY , the variance in Y , to calculate a coefficient of determination, D using formula 7.11.

$$D = \frac{VY - RV}{VY} \quad (7.11)$$

$\sqrt{D}$ is referred to as the multiple correlation coefficient. In the second stage of NORMKA's operations a new function of type $Y = A0 + A1X1 + A2X2$ is determined using new values for $A0$ and $A1$. RV and D are then recalculated. This procedure is repeated for all the T dependent variables. NORMKA handles the variables in the order they are given without any attention being paid to each variables relative importance. The NORMKA program is therefore less advanced than the multiple regression program BMD02R (Dixon 1968).

The results from NORMKA are given in the form of an optional detailed print out from the line printer and are also stored in a 1-dimensional field RES . The resulting regression coefficients are given with correction made for the transformations carried out on the original data.

7.7.1 The level of accuracy of NORMKA

In the section on numerical terrain models it was pointed out that a mathematically defined surface could be fitted to a collection of points obtained by a survey. This is possible as every point can be described by a numerical triple of type (X,Y,Z) , where X and Y refer to the coordinates of a point and Z to the value of the function at that point. The determination of this surface can be carried out with the method of least squares and the NORMKA procedure. In this case Z is taken to be the dependent variable while X and Y are taken to be independent variables. If a special input subroutine is used more complicated surfaces than the simple plane $Z = A + BX + CY$ can also be determined. Use of this subroutine allows powers and products of X and Y to be included in the equation. If this is done care must be taken to avoid truncation errors. As variables in the NORMKA procedure are of type double precision numbers up to 18 figures long can be catered for when using the UNIVAC 1108 computer. Thus if X is a 3-figure number X^3 can be used as an independent variable without truncation errors occurring.

7.8 Trend surface analysis

A surface fitted to a set of points in 3-dimensional space can be called a trend surface and the actual calculation of the surface trend surface analysis. This method of analysis was originally developed by geologists, but it can of course be used in conjunction with functions other than height over sea level. If one is studying a phenomena with a geographical distribution, for example an infectious disease, the trend surface analysis is possibly of interest. The date of the first occurrence of a particular illness at point (X,Y) can be referred to as Z . The spread of the plague during the 13th century has been studied in this manner.

Trend surface analysis is also applicable when studying different types of innovations. Z can in this case refer to the point in time when an individual in the vicinity of point (X,Y) adopted a new innovation. As an example of this use of trend surface analysis in innovation studies one can mention a study of the spread of opium dens in India.

A further example of the movement of a phenomenon across a surface is that of the coming and going of the seasons of the year together with associated effects. The date of the arrival of the swallow has been used here to illustrate the application of this technique. In 1970 the swallow arrived in southern Sweden in the middle of April and reached the north of the country at the end of May. If the day of arrival, counted from the first of April without regard for the month, is referred to as Z then a swallow arriving on the 21st May has the Z value 51. X and Y are as usual the coordinates of a point, using the coordinate system of the land use map. The unit used for the coordinates is in this case 10 kilometres. A trend surface analysis using this data gives as the resulting equation $Z = -26 - 0.0428X + 0.114Y$.

The deviation of the observed values from this plane is large which gives a multiple correlation coefficient with the low value of 0.42. A further result of this study was that one obtains a multiple correlation coefficient with nearly the same value when only Y is used as independent variable. The correlation coefficient becomes larger when X^2 , Y^2 and $X \cdot Y$ are used as independent variables, but this effect is proportionally very small and a satisfactory result can be obtained with the simple equation $Z = -25 + 0.103Y$.

7.9 Coordinate transformation

The transformation of coordinate values from a provisional coordinate system of a map to another system occurs fairly often. The usual approach

is to reconstruct the desired coordinate system and to draw it out on the map. This coordinate system is then used when determining the coordinates of the points in question. Unfortunately factors, such as distortion of the paper and the projection used, combined with the problem of reading off the new values make this approach inaccurate. The measuring of the coordinates in terms of a provisional coordinate system is simpler to carry out, although, the problem of coordinate conversion of course remains. The provisional coordinate system is referred to as the (xk,yk) system in the following discussion, and the desired coordinate system is referred to as the (x,y) system. If the coordinates of three points are known in both system the relationship between the two systems can be determined. This is a simple affine transformation from the one system to the other as shown by equations 7.12.

$$\begin{aligned}x &= a1 + b1xk + c1yk \\ y &= a2 + b2xk + c2yk\end{aligned}\tag{7.12}$$

The problem is thus to determine the 6 constants in these two equations. When this has been done the coordinates of the points can be determined in terms of the desired coordinate system.

A drawback of the simple affine transformation is that it is not possible to discover if a point has been measured inaccurately. Large errors can therefore occur in the calculated (x,y) values. In order to reduce the error risk with the above method the constants are calculated using more than three points, the coordinates of which are known in both systems. This is done with the method of least-square and the NORMKA procedure. Use of this approach also makes possible the use of more complicated functions than the simple planes given by equations 7.12.

7.9.1 The KOGRAF program

The subroutine which calculates the constants for use in the coordinate transformation equations and which also performs the actual transformations is called KOGRAF. This subroutine is particularly suitable when transferring information from a map, drawing or aerial photograph to coordinate based data for further processing with the NORMAP system.

8 Comparison of isarithmic maps

The problem of carrying out a comparison between two isarithmic maps occurs fairly often. This is often carried out visually and in certain cases it is possible to detect whether or not there is a similarity. In the majority of cases this is not possible, due to the effect of different shading methods, different isarithm density and different isarithm values. A visual comparison between two isarithmic maps can be aided by the overlaying of one set of isarithms over the other, but an even more precise comparison becomes possible when a numerical approach is adopted.

8.1 Isarithm overlay

This technique is based on the selection of an isarithm from one of the maps and the entering of this isarithm in its correct position on the other map. This procedure can of course be carried out for more than one isarithm. This method is extremely simple but also effective, especially when the two maps have a similar character. Figures 8.1 and 8.2 are two isarithmic maps which are to be compared. The maps show the distribution of population in Kronoberg county in 1950 (figure 8.1) and in 1960 (figure 8.2). The reference area is in this case a square with sides 10 kilometres. The original data has been obtained from Claeson's population square net maps (Claeson 1964, Claeson & Stanisic 1969). A direct visual comparison of the two maps is of course possible but presents practical difficulties. In this case the overlay technique has been used and the result for the 1000 isarithm is shown in figure 8.3. The result shows how the population of the rural districts has declined in the years between 1950 and 1960. A prerequisite for this type of comparison is access to two correct and comparable maps, which can best be arranged by using the NORMAP system in conjunction with coordinate based population data.

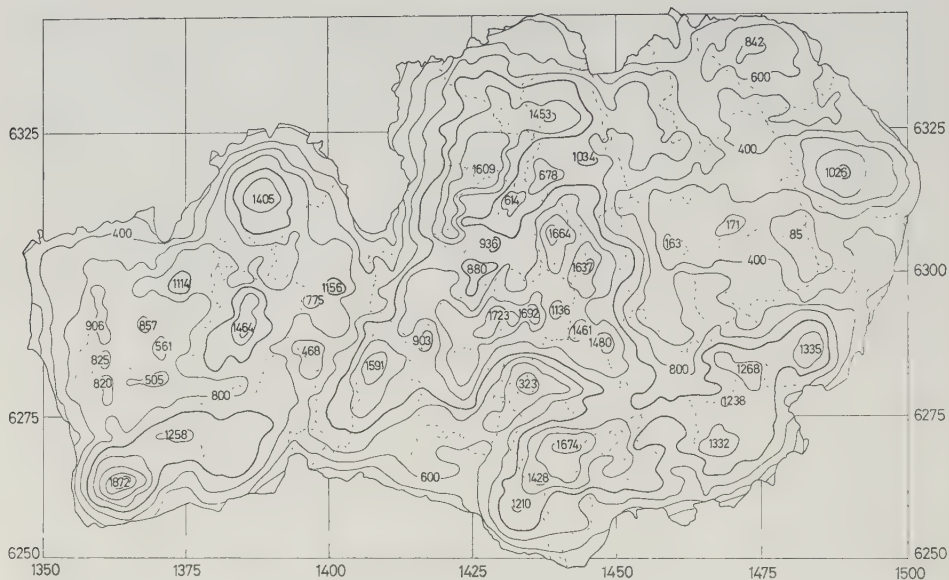


Figure 8.1. Isarithmic map showing the distribution of rural population in Kronoberg county in 1950. Reference area: 100 square kilometres. Data: Claeson & Stanisic 1969.

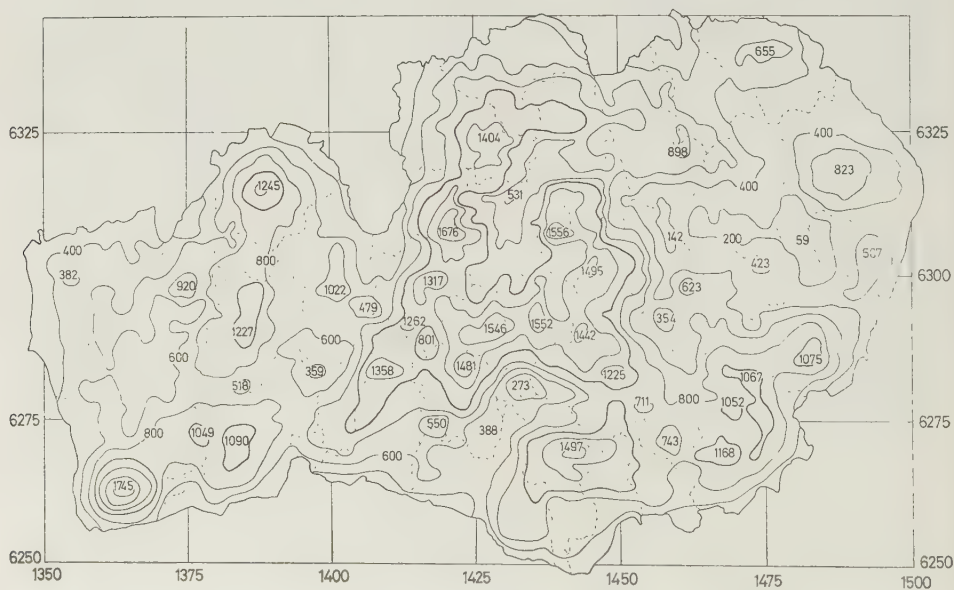


Figure 8.2. Isarithmic map showing the distribution of rural population in Kronoberg county in 1960. Reference area: 100 square kilometres. Data: Claeson 1964.

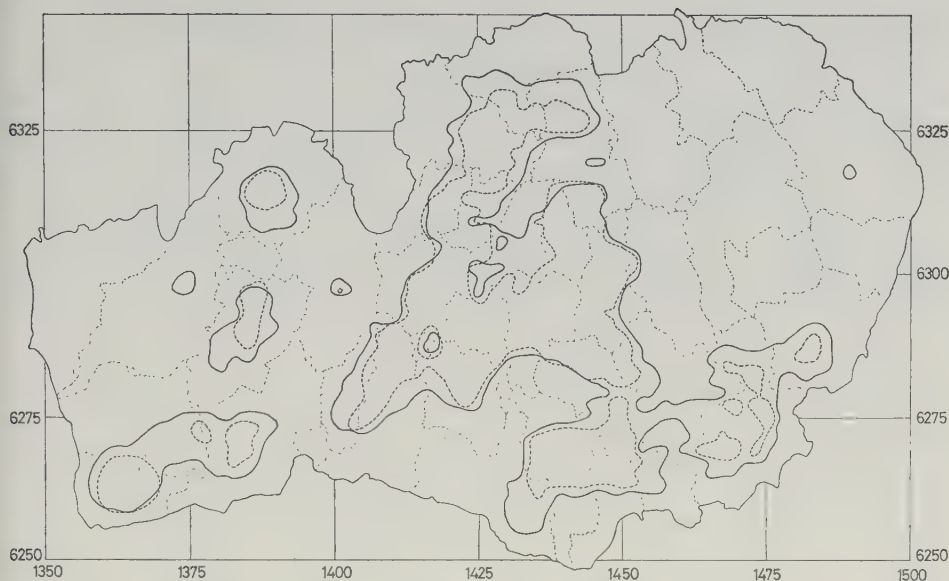


Figure 8.3. Map with overlaid isarithms. The 1000 isarithms have been copied from the maps shown in figures 8.1 (solid lines) and 8.2 (dashed lines).

8.2 Comparison between isarithmic maps using the quotient and difference methods

Numerical comparison of isarithmic maps can be carried out using three methods; the difference method, the residual method and the correlation method. All of these methods are also applicable to other types of numerical map.

All of the above methods have as their basis the fact that the value of the function is known for the same collection of grid points on both maps. If the value of the function at a point on map 1 is known as X and the corresponding value for map 2 is Y the quotient of these two values X/Y provides one way of describing the variation between the two maps at that point. Figure 8.4 is such a quotient map prepared in the above manner. This simple method is only applicable when the two variables are directly proportional to one another, i.e. $Y = AX$. If one instead believes that the relationship between the two variables is of type $Y = A + X$ the difference between X and Y should be used.

8.3 The residual method

The difference and quotient methods can be developed further if the assumption is made that the same changes have occurred throughout the

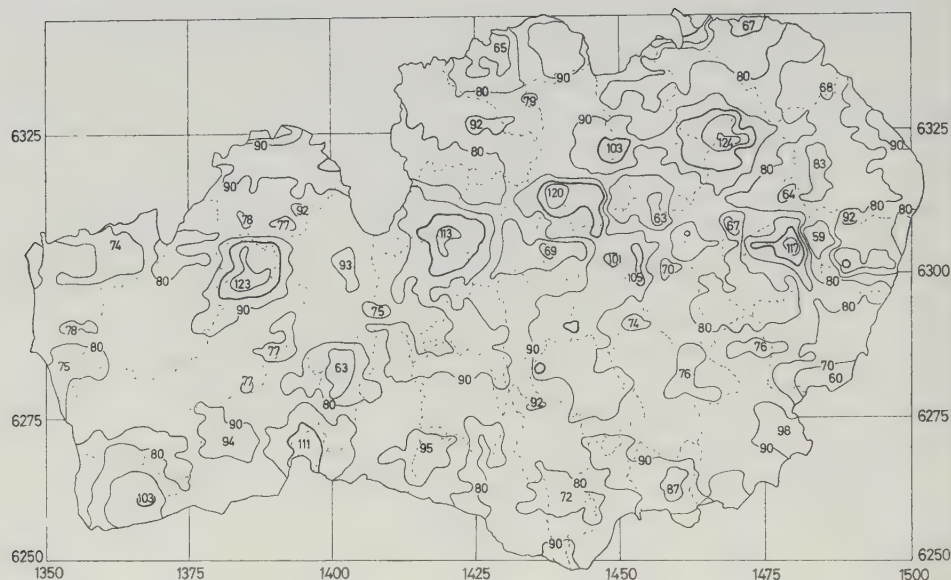


Figure 8.4. A quotient map of Kronoberg county. The function mapped is the population in 1960 expressed as a percentage of the population in 1950. The 1960 population is as shown in figure 8.2 and the 1950 population as shown in figure 8.1.

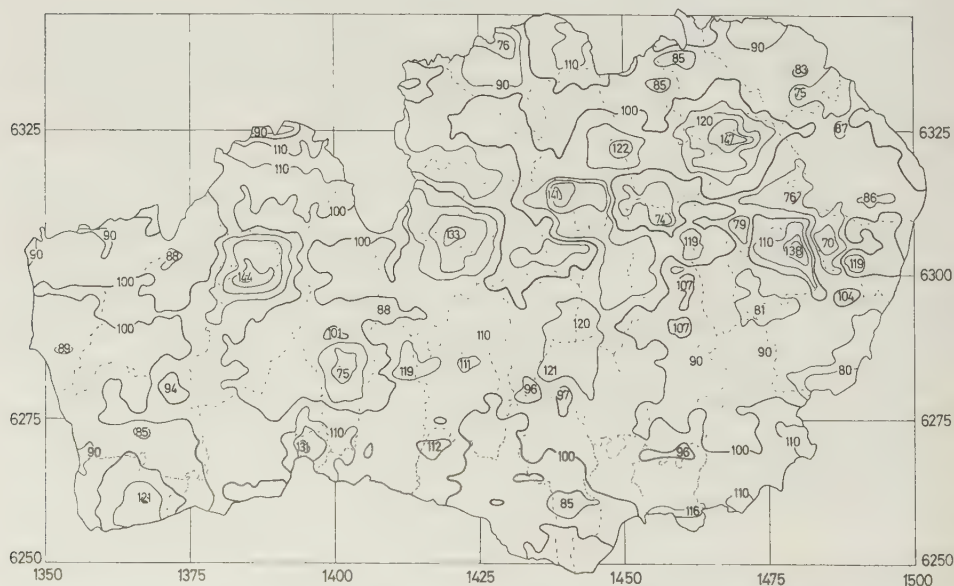


Figure 8.5. An isarithmic map showing the 1960 population as a percentage of the theoretical population, i.e. that population the rural areas would have had if population decrease during the 1950's had been evenly distributed throughout the county.

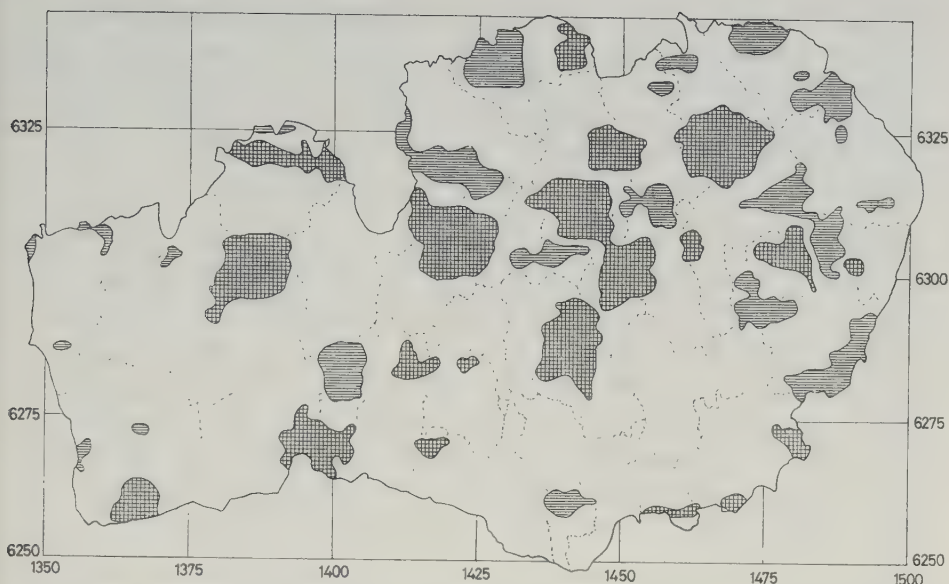


Figure 8.6. A residual map based on the map shown in figure 8.5. Areas where population decrease has occurred faster than expected (areas with negative residuals) are shaded horizontally. Areas where population decrease was less than expected or where population increase occurred (areas with positive residuals) are shown cross shaded.

area. Thus for example we know that the population of the country areas of Kronoberg county decreased by 20 % during the 1950's. Thus if the population in year 1950 is referred to as X and the population 1960 as Y the change between the two dates for the whole county can be represented by the equation $Y = 0.8X$. If the same percentage change had occurred throughout the county then the above relationship is true for every part of it. Thus Y becomes a theoretical value with which the actual value for year 1960 can be compared. The most satisfactory way of presenting this comparison is to give the population in 1960 as a percentage of the theoretical value. This percentage can then be mapped in the form of an isarithmic map in the normal manner, see figure 8.5. A certain amount of deviation from the calculated value is of course of no great importance, perhaps up to 10 %. If this 10 % rule is adopted three areas on the map present themselves. The first where the decrease in population is larger than expected. These are usually referred to as areas with negative residuals. The second area is that where the change in population is approximately the same as that for the county as a whole. The third area covers those parts of the county where the decrease in population has been less than expected or where an increase in population has occurred.

This third area is usually said to have positive residuals. Figure 8.6 is thus a simple residual map showing the areas with negative residuals hatched horizontally and those with positive residuals crosshatched. One can conclude from the above map that the positive residuals lie in the vicinity of such towns as Alvesta, Älmhult, Ljungby and Markaryd. On the other hand Växjö lies in the area with normal percentage change.

In most cases simple models, such as $Y = X - A$ or $Y = AX$ are felt to be insufficient and more complicated functions are therefore applied. The problem contains two variables X and Y and can therefore be treated as a curve fitting problem.

The first stage is to enter each point into a coordinate system having X and Y coordinates, and the next is to decide what type of function is likely to give the best fit. A straight line ($Y = A + BX$) can suffice, but often it is necessary to use higher degree curves. Once this decision is made the method of least-square can be used to calculate the coefficients.

The value used when preparing the new map is the deviation between the observed and the calculated value for Y . This latter value for Y is calculated by setting in the corresponding X value in the obtained function. These deviations are usually standardised by division with the standard error of estimate, i.e. the deviations are given units of size 1 standard error of estimate.

When the absolute value of the deviation between the calculated value of Y and the observed value is less than 1 standard error of estimate the situation is accepted as normal and the point assigned to the normal area. The normal area in this case covers about 2/3 of the map. The area with positive residuals with deviation greater than 1 standard error of estimate, covers approximately 1/6 of the map as does the area with negative residuals. These proportions only apply if X and Y are normally distributed. Other boundaries for the different areas can of course be used and the map divided into more than three areas.

The major advantage with the residual method is that X and Y do not need to be of similar character; the only requirement is that they are correlated. Thus the population of rural areas Y , can be explained using the degree of cultivation X . In Kronoberg county the correlation between the two variables X and Y in 1960 was 0.71. There is thus a clear relationship between the isarithmic maps showing population distribution and the distribution of cultivated areas. This relationship can be described by the equation $Y = 63.2 + 0.114X$, with Y in this case equal to the number of residents in rural areas per 25 km² and X equal to the number of hectares of cultivated land occurring within the same reference area.

8.3.1 The NORRES procedure

NORMAP contains a procedure NORRES for the carrying out of comparisons between two numerical maps using the residual method. The values to be compared are referred to as X and Y . A prerequisite for the operation is that the two maps cover precisely the same area, thus having the same origin and the same values for L (the number of rows) and W (the number of columns). The area to be mapped forms a polygon which lies completely within the rectangle formed by L and W . In most cases the polygon is smaller than this rectangle and points in these intervening areas are given the value 0. Such points are of course not used when calculating the residuals. The presence inside the polygon of a point having a 0 value can be detected by use of the point-in-polygon procedure NORP in conjunction with the coordinates for the vertices of the polygon. Use of NORP for this purpose is on the other hand extremely uneconomic and a better approach is provided by using the rule that points having zero values for both X and Y are not included in the calculation. When points of this type occur inside the polygon containing the area to be mapped they are given a very small X or Y value when the data is prepared and are thus included in the calculation.

The parameter T refers to the degree of the function which is fitted to the collection of points (X,Y) . The procedure works in stages, i.e. it begins by calculating the residuals using a straight line $Y = A_0 + A_1X$. If T is greater than 1 the procedure carries on to the next stage where a second degree curve is used, $Y = A_0 + A_1X + A_2X^2$. In the third stage a third degree curve is used and so on until the final curve of degree T has been obtained. The values $X, X^2, X^3, \dots, X^T$ can become exceedingly large and it is therefore necessary to transform the original X -values.

The subroutine NORMKA is used to carry out the actual curve fitting. As the user of the NORRES procedure is concerned only with residuals the data does not require transformation back to the original values.

The result from NORMKA is printed out on the line printer in conjunction with the execution of NORRES while the resulting deviations from the polynomial of degree T are placed in *MAPX*, the same results are placed in *MAPY*, but in this case divided by the standard error of estimate and multiplied by 100. The contents of both the above arrays are rounded off to the nearest integer.

A modified version of the NORRES procedure makes possible multiple residual analysis. This means that comparison can be carried out between more than two maps. If, for example, 4 maps are to be compared, showing

the variables Y , X_1 , X_2 and X_3 then the following functions could be tried in order, $Y = A_0 + A_1X_1$, $Y = A_0 + A_1X_1 + A_2X_2$ and $Y = A_0 + A_1X_1 + A_2X_2 + A_3X_3$. There is of course no obstruction to the use of polynomials of higher degree than 1 when carrying out this type of multiple residual calculation.

8.4 The correlation method

When carrying out a comparison between two isarithmic maps using the residual method the correlation between the two variables X and Y is calculated. When carrying out this operation all the pairs of values lying inside the area of the map were used and the result referred to the whole area of the map. The correlation coefficient for sub-areas of the map can of course be calculated in a similar manner (Robinson 1962). In this way variations in the correlation coefficient over the area of the map can be detected.

A computer is of great help when comparing two maps using the correlation method but in this case the approach used must reflect the computer's method of operation. The best way of doing this is to calculate the correlation at every grid point. This is done by treating the grid point as the centre point in a reference square with sides of length S . The reference square thus covers S rows and S columns.

When the regular grid from the NORI program is used the grid points lie in zig-zag lines, but every row and every column contains exactly S grid points.

Overlapping reference areas are used when carrying out this operation instead of the method used by Robinson (1962). The result of these operations is a new numerical map showing the same number of values as the original maps. This new map can then be processed further to produce an isarithmic map. Figure 8.7 is a map of this type showing the correlation between population density in country areas (number of persons living in country areas per 25 km² reference square), and the degree of cultivation (the percentage of the reference area devoted to cultivation). The original map showing the degree of cultivation has been obtained from Abrahamson (1963), the population distribution has been obtained from Claeson's population map of Kronoberg county (1964). In the case of the whole county the correlation between these two variables is about 0.7. As can be seen from the map, figure 8.7, there are large local deviations from this value. There exists several areas with a higher correla-

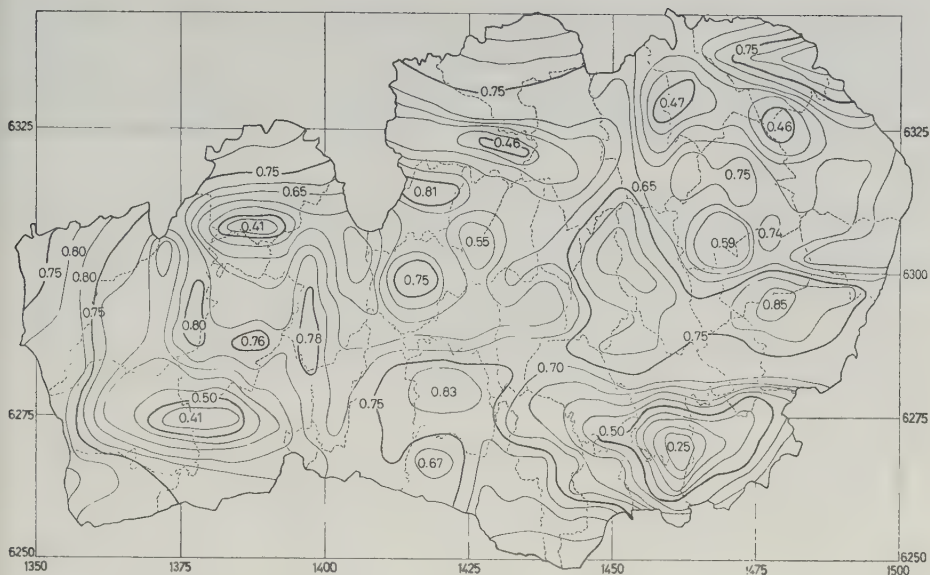


Figure 8.7. Isarithmic map showing the correlation between population density in country areas and the degree of cultivation. Reference area: 25 square kilometres. Number of grid points used when calculating correlation: 25. Data: Abrahamson 1963, Claeson 1964.

tion than 0.80, together with several areas with a lower correlation than 0.50. The lowest coefficient is 0.25 while the highest is 0.85.

The procedure used for comparison of two isarithmic maps using the correlation method is called NORCOR. The completed map is stored grid point for grid point in their order of calculation. When all grid points have been treated they are transferred back to *MAPX* after multiplication with 1000.

8.5 Auto-correlation

8.5.1 Time series

Auto-correlation is the comparison of a set of data with itself. As an example one can use a series of items of meteorological data T_i which one suspects is periodical in nature. The problem is then to discover the length of the period and the points at which they begin. This can be done by copying the original information to give a new set of information N_i . The correlation between these two bodies of information is of course 1.

The elements in N_i are then moved one position, i.e. item N_1 is compared with item T_2 and so on. In this new situation the coefficient of correlation becomes less than 1. The procedure is repeated until N_1 corresponds with T_n where n is the number of items of data minus a freely chosen number, for example 25.

The result of the above operations is a series of auto-correlation coefficients which are periodical only when the original data is periodical. The first coefficient of correlation in the series is 1 or perhaps a tiny fraction less, as is the coefficient lying at the beginning or end of every new period. Those coefficients which lie in the middle of a period are on the other hand equal to or a fraction larger than -1 .

In the case of meteorological information T_i can refer to the number of seconds of daylight occurring on day i . An auto-correlation analysis of this information would show that the correlation coefficients were approximately 0 around days 91 and 273. A strong negative correlation is present around day 182 while complete correlation would occur around day 365.

Daylight is an uncomplicated variable having only one period, the year. Air temperature on the other hand is a more complicated variable, which varies not only from day to day but also from year to year. The period of daylight on a certain day varies very little from year to year but this is not true of air temperature. Thus an auto-correlation analysis of air temperature does not give such a clear result as is the case with length of daylight. It is thus highly unlikely that the correlation coefficients are exactly 1 or -1 . Therefore it is probably necessary to graph the coefficients and to use the maximum and minimum values to determine the length of the shorter periods. The whole process can then be repeated using one value from each period, perhaps the first, in order to determine if longer periods exist. This technique is only applicable when the longer period is approximately a multiple of the shorter period which can often be the case when the longer period is considerably longer than the shorter period.

8.5.2 Dispositions in space

The technique described above can easily be extended to cater for 2-dimensional variables and thus to geographical problems. The original data can be a population map in numerical form which is then moved one cell at a time over its copy, the comparison is carried out for every cell in every row and for every row in the map. The result of this operation is a new numerical map containing coefficients of correlation.

8.5.3 The auto-correlation program NORAUT

The auto-correlation program NORAUT can be used when analysing both 1- and 2-dimensional distributions. 1-dimensional data is treated as if it were a map with only one row.

The numerical map used for this operation can have been read in by using MAPIN or constructed with help of one of the mapping programs, for example NORK, NORKO or NORI. Thus NORAUT is not limited to a rectangular grid net but can also handle triangular grids.

When calculating coefficients of correlation it is necessary to use a sufficiently large number of data points. NORAUT does not calculate the correlation coefficient at a grid point if less than 25 values are available. The input data to NORAUT is stored in the array *MAP*. As this array actually forms a rectangle around the actual area to be mapped it will contain grid points having the value 0 i.e. they lie outside the area of study. These points are not included in the correlation calculations.

The values stored in *MAP* should always be integer. If one has decimal values they should be multiplied by a suitable power of 10 to give integer values. This operation does not affect the correlation calculations in any way.

The correlation coefficients obtained are multiplied by 1000 and then rounded off to the nearest integer. They are then placed in an array *MAPC* which can be printed out on the line printer using MAPOUT. If *MAPC* is to be used as input to NORIP care must be taken to see that it does not contain negative values. If so is the case the map can be transformed through use of MAPTRF to give results lying between 0 and 2000 which are suitable for use with NORIP.

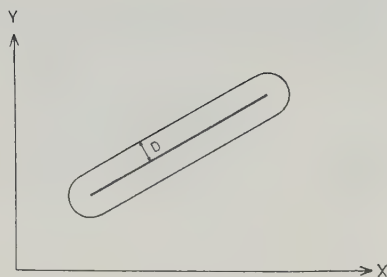
9 Mapping of traffic flows

The NORI program is used to map the population within a given reference area. If a circle is used as the reference area an alternative definition of the mapped function can be used, i.e. the number of persons living within a certain distance of the grid point. When dealing with traffic and traffic flows a similar function can be used i.e. the number of trips which pass within a certain distance of a point. This function exists at all points including grid points. One way of preparing this type of map would be to place a reference area of some kind around the route taken by a particular journey and to add 1 to the value of the function at every grid point which lies inside that reference area. A prerequisite for this approach is that the exact route is known in terms of which sections of the road network are used for a particular journey.

9.1 Straight line journeies

The simplest assumption that can be made about the route taken between two points is that it forms a straight line; this is the route "as the crow flies". This route cannot be followed if impassable barriers lie across it and is therefore often referred to as the trip desire line. This type of journey can be mapped as a collection of straight lines between the different starting and goal points. Unfortunately this type of map is difficult to read and analyse. Several attempts have been made to improve this type of map without much success, probably due to the fact that no rigorously defined function is used. Transfer of one of the existing manual methods to a computer program appears highly questionable due to the fact that the resulting program would be slow and uneconomic. A new method suited to the computer's methods of operation is therefore required. The method used in NORMAP uses a function defined as the number of journeies which pass within a given distance of each point. The resulting values at each grid point form the basis of the map.

Figure 9.1. Every point within the reference area shown in the figure lies nearer the straight line than the distance D . The reference area consists of a rectangle and two half circles.



9.2 Linear journeys and the NORILK procedure

The NORILK procedure maps straight line journeys. A similar grid net to that of NORI is used and the result is a numerical map which can be processed further to give an isarithmic map by use of the NORIP procedure. As is shown in figure 9.1 the reference area used in NORILK is a rectangle with a half circle at each end. The width of the rectangle is $2D$ as is the diameter of the two half circles. The procedure NORILK uses only the starting point (SX,SY) and the goal point (GX,GY) to describe the journey and pays no attention to the actual road network. In certain cases journeys must cross a barrier of some type which can only be crossed at certain points. This problem is highly complicated and is not therefore included in this discussion.

Figure 9.2 is an isarithmic map showing the distribution of journeys in the vicinity of the school Vegaskolan in Lund. The journeys shown are only those of school-children on their way to school. The map has been prepared with help of the NORILK procedure and clearly shows which routes have the highest frequency of use. It would appear desirable that these routes did not cross routes with heavy traffic. Thus, a map of the type shown in 9.2 can be used when deciding where underpasses and footbridges should be located in order to bring about safer journeys to school.

9.3 Journeys via actual road networks and the NORILK procedure

Although the procedure NORILK makes possible the mapping of journey frequency one has the feeling that the actual road network plays an important part. In order to check this the actual network must be available, preferably in the form of coordinate based nodes and links. The function to be mapped is the same, i.e. the number of journeys which pass within a given distance of each point. Every route is described in the form of a



Figure 9.2. Pedestrian traffic created by children on their way to Vega School in Lund on the 26th of May 1966. All journeys given as straight lines.

polygon train and surrounded by a reference area built up of rectangles and circular sections. See figure 9.3. Every vertex in the polygon train, including the start and goal points, is made the centre of a circle having a radius D . Every link in the polygon train is placed on the axis of a rectangle of width $2D$ having the same length as the link in question and used independently as input to NORILK. One effect of this approach is that certain grid points are included twice and correction must be made for this inaccuracy. This correction is carried out by giving every grid point falling inside the reference area a number. This number can be of the type, $(C - 1) \cdot L + R$, where R is the row number of the point and C the column number. L is as usual the length of the map, i.e. the number of rows it

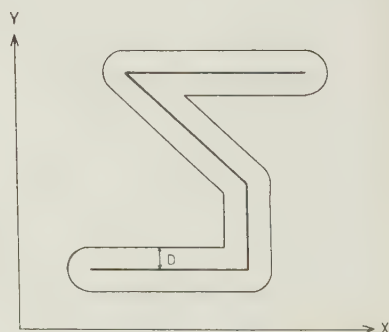


Figure 9.3. Every point within the reference area shown in the figure lies nearer the zig-zag line than the distance D . The reference area consists of a number of rectangles and segments of circles.

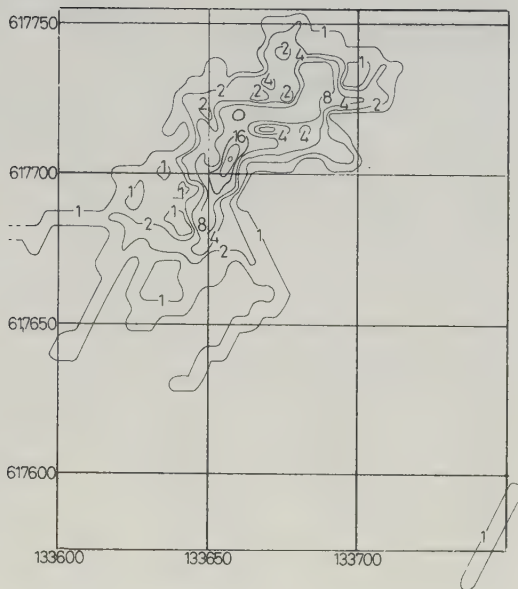


Figure 9.4. Pedestrian traffic created by children on their way to Vega School in Lund on the 26th of May 1966. Actual routes.

contains. When every link has been treated the resulting grid point numbers are arranged in ascending order and checked to see that every number only occurs once. The unrequired grid points are removed from the list and the value of those remaining increased by 1. The procedure based on this approach is called NORIRK.

Figure 9.4 is an isarithmic map showing the distribution of the same journeys to school as were illustrated in figure 9.2. In this case NORIRK has been used in conjunction with a description of the actual routes taken. A comparison of these two maps is exceedingly interesting, as visually they are similar.

This type of information about the actual route taken is only rarely available and one can therefore draw the conclusion that the procedure has only limited application. The lack of this information can be made good by the use of one of the large range of traffic distribution models which exist. These models choose a route between a starting point and a goal point via a given traffic network. The best known of these is probably the shortest route model. In this model the assumption is made that a journey always follows the shortest route between the two points. This problem is discussed further in chapter 13 of this report.

10 Range maps

The basis of the NORI procedure is the use of reference areas of constant size. NORI can thus be used to map population density, i.e. the number of persons living in a given reference area. On the other hand when a circle is used instead of a square as the reference area the function can be defined in a somewhat different way. In this case the resulting population is also the number of persons living within a certain distance of the actual grid point. In chapter 6 it was shown that populations could be treated as volumes. One can therefore call the population of a reference area a reference volume. The two terms reference area and reference volume can be used to explain how NORI functions, i.e. the reference area is constant and the reference volume varies. Using these two terms a further function can be defined, the range function, where the reference volume is constant and the reference area varies in size. A map of this function shows population range, usually in the form of the length of radius of the reference circle.

10.1 Constant reference volume, dot maps

The use of a constant reference volume is not new. It occurs in connection with population dot maps; the oldest maps of this type are more than 100 years old (de Montizon 1830, Mentzer 1859). The construction of a dot map is begun by dividing up the area to be mapped into a number of reference polygons, each of them containing a predetermined number of persons. Every polygon is then represented by a dot at its centre. Sidenblad (1880) refined the reference polygon method when calculating the "mean living space" (*medelegovidden*), for the different Swedish counties. A dot map is characterised by the fact that the reference polygons cover the whole of the mapped area without overlapping. It is quite possible to develop an algorithm for the production of dot maps (Olsson & Selander 1970). On the other hand a dot map is a fairly primitive type of map from which it is difficult to obtain numerical information. The value of a program of this type is thus highly questionable.

10.2 Range calculation algorithms

When the reference areas are allowed to overlap one another the resulting program is considerably easier to develop than is the case with a dot map program. Information in the form of a square net map is a prerequisite when developing this type of map. Three methods can be used when calculating the population range, *RANGE*, at a point. The simplest of these methods involves the calculation of the distance between the grid point and every data point, i.e. the centre of every cell in the map. The distances are then sorted into ascending order and summed from 0 upwards until the constant reference volume is reached. The distance from the grid point to the last data point used in this operation is the population range at that grid point. This procedure is then repeated for every grid point. A somewhat faster variant is to allocate the different distances to classes, the population of every class of distances is calculated at the same time as the distances are calculated. The population of the first N classes is then summed. The value of N is such that the sum of the first N classes is equal to the reference volume. The range is thus N times the class width. Of course in most cases the sum is not exactly equal to the reference volume, in such cases N is given a value such that the sum of the first N classes is less than the reference volume and the sum of the first $N + 1$ classes is greater than the reference volume. A more accurate value for *RANGE* is then obtained by linear interpolation between N and $N + 1$.

The range map construction methods described above are very simple to program but very expensive to use. They should therefore only be applied in special cases where the reference volume is more than 50 % of the total population of the mapped area. If barriers are present in the mapped area then the methods described above are probably the most applicable.

A second approach to the construction of range maps also exists. With this method a small reference area having a diameter of 2 units is placed centrally over a grid point and the population calculated. If this population is equal to that of the reference volume the range at that grid point is 1 unit. If the population of the reference area is larger than the reference volume a more accurate range can be calculated by interpolation between 0 and 1. On the other hand when the population of the reference area is less than the reference volume its diameter is increased by 2. The new reference area is then treated in the same way as the original. This procedure is repeated until a reference area having a population larger than or equal to the reference volume is finally obtained, thus allowing an approximate value for *RANGE* to be determined by linear interpolation. This operation is repeated for all the grid points making up the map. This

second method should be used when the reference volume is less than 50 % of the volume of the map (Nordbeck & Rystedt 1969a, 1972).

In cases where the reference volume is very small the third method can be used. This involves repeated use of the NORI procedure in a similar way to that described in section 5.1. The result of each of these three methods is a population range map in numerical form. This can then be used as input to NORIP in order to give an isarithmic map.

Programming the distance class approach is straightforward and will not be described further in this report. The NORI procedure was described in chapter 4 and therefore the rest of this description will be concentrated on the approach using an expanding reference area.

Unfortunately all three methods lead to very slow computer programs. It requires many hundreds times as much computer time when constructing a population range map as when using the NORI procedure in the calculation of a population density map. Great care must therefore be taken when using these techniques, in order to avoid exceedingly high costs. Furthermore the method used should be chosen with regard to their different characteristics.

10.3 The reference area's shape

The population range at point (X,Y) can be defined as the distance it is necessary to travel from that point in order to reach a given population. Another description of population range is the radius of a reference circle containing a given population. This reference circle has point (X,Y) at its centre. This definition of population range implies a circular reference area which unfortunately complicates the necessary calculations. It is therefore not unreasonable to wonder if the circle could not be replaced by a square. Use of a square produces a somewhat inaccurate range map but this can to a great extent be compensated for. The variable SB is given a value equal to half the length of side of the reference square, but is in this case a little too short when compared with the accurate value of $RANGE$. This inaccuracy can be corrected by multiplying SB with a suitable constant. A value for this constant of $2/\sqrt{\pi}$, approximately 1.13 is the most satisfactory given that the population is evenly distributed over the surface.

Range maps with a constant reference volume find their main use when calculating transport costs in conjunction with the siting of facilities requiring a certain customer base. This problem is discussed further in section 10.6. It is therefore of interest to study how these costs are affected

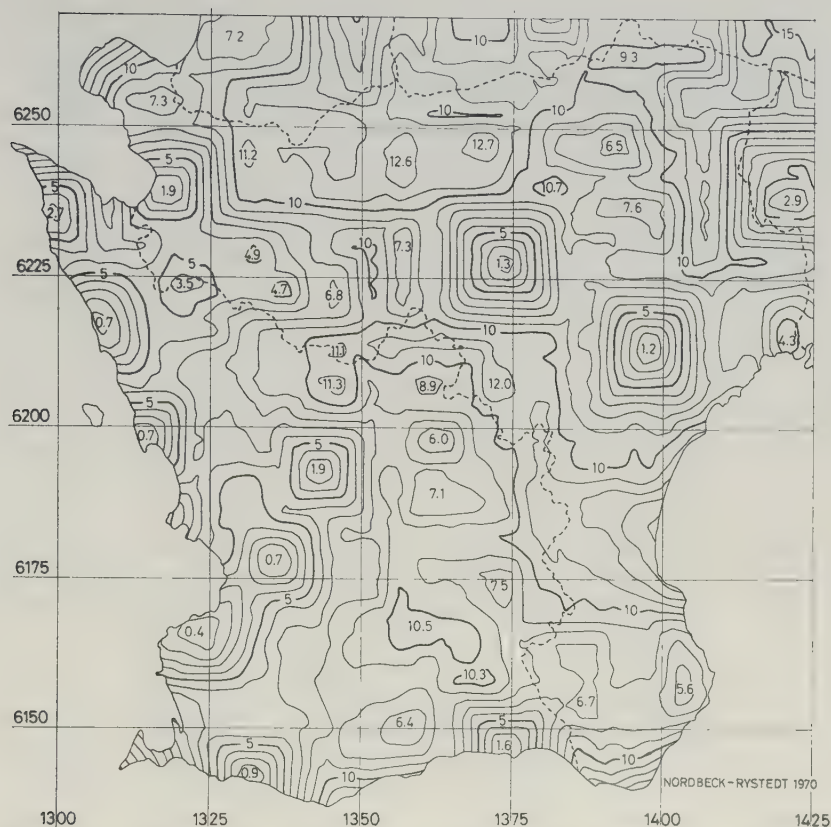


Figure 10.1. Range map of Skåne and neighbouring areas of Halland, Småland and Blekinge. The map shows half the length of side of the reference square. It can be seen from the isarithms that a reference square of varying size has been used instead of a reference circle. The isarithms form nests of squares in the vicinity of the larger built-up areas. Reference volume: 10 000 persons.

when a reference square is used as an alternative to a reference circle. A suitable measure of the degree of change is provided by the mean distance between points evenly distributed over the surface and the centre of the circle or square. Use of formula 10.4 gives a mean value for a square circumscribed a circle about 15 % larger than the mean value for the circle.

When a square having the same area as the circle is used the mean distance is approximately 1.5 % larger. This inaccuracy is small enough not to need further correction.

The correlation between the two variables *SB* and *RANGE* is very high. Only in the vicinity of larger towns where population density shows larger variations do small inaccuracies occur when using $RANGE = 1.13SB$.

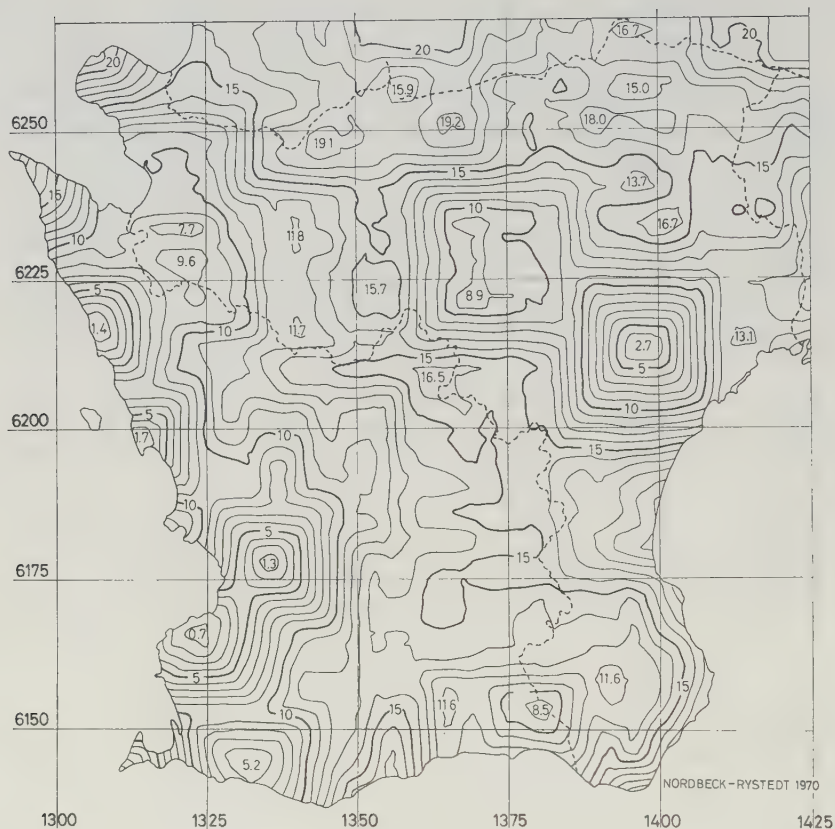


Figure 10.2. Range map of Skåne. Reference volume: 25 000 persons.

These inaccuracies do not prohibit the use of a reference square when constructing population range maps especially as the resulting program is considerably faster than the corresponding program using a reference circle.

10.4 Range maps of Skåne

Figures 10.1—10.4 show a collection of range maps of Skåne (Scania). One can see from these maps that the population range is 15 to 20 times longer in the northeastern part of the region than in the area around Malmö.

A square reference area was used when constructing these maps as can be seen from the shape and position of the isarithms. In the vicinity of larger towns the isarithms have a somewhat rectangular form and lie parallel

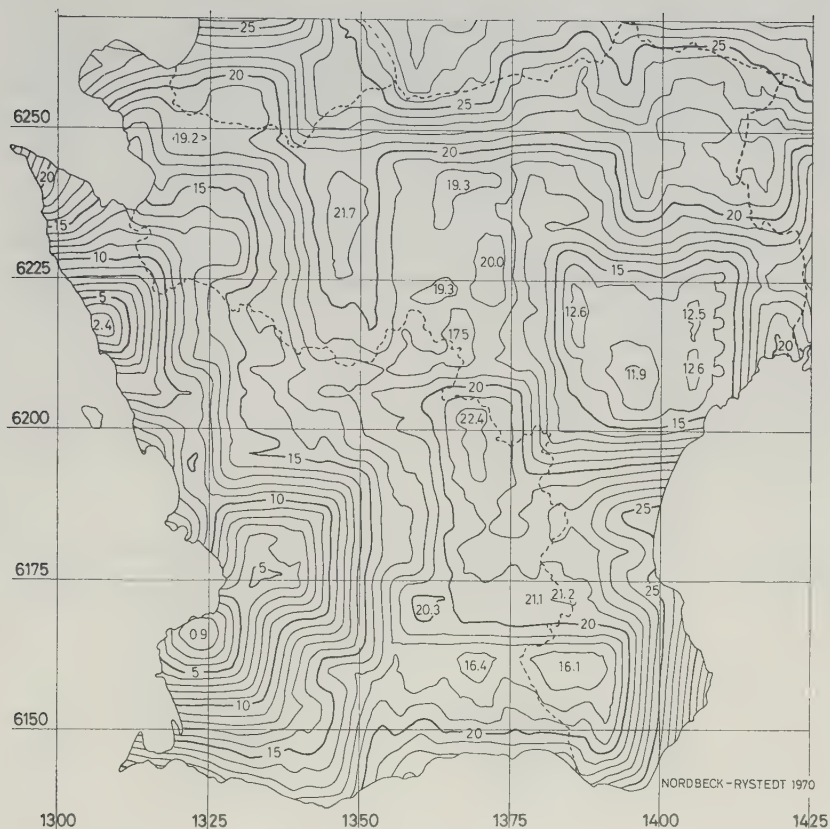


Figure 10.3. Range map of Skåne. Reference volume: 50 000 persons.

with one another. If reference circles had been used the resulting isarithms would have been built up from parallel circular sections. The grid net equidistance in the case of all 4 maps is 2 km. Even when using such a small equidistance it is possible to detect the form of the grid net from distortions in the north—south isarithms in the northeast corner of the region. If a coarser grid net had been used this effect would have been even more apparent.

10.5 Fraction maps

The proportion of a given population which satisfies a given requirement is a function which is often mapped. A proportion of this kind is often referred to as a fraction. Thus the corresponding map can be called a fraction map. Well known fractions are the proportion of working persons

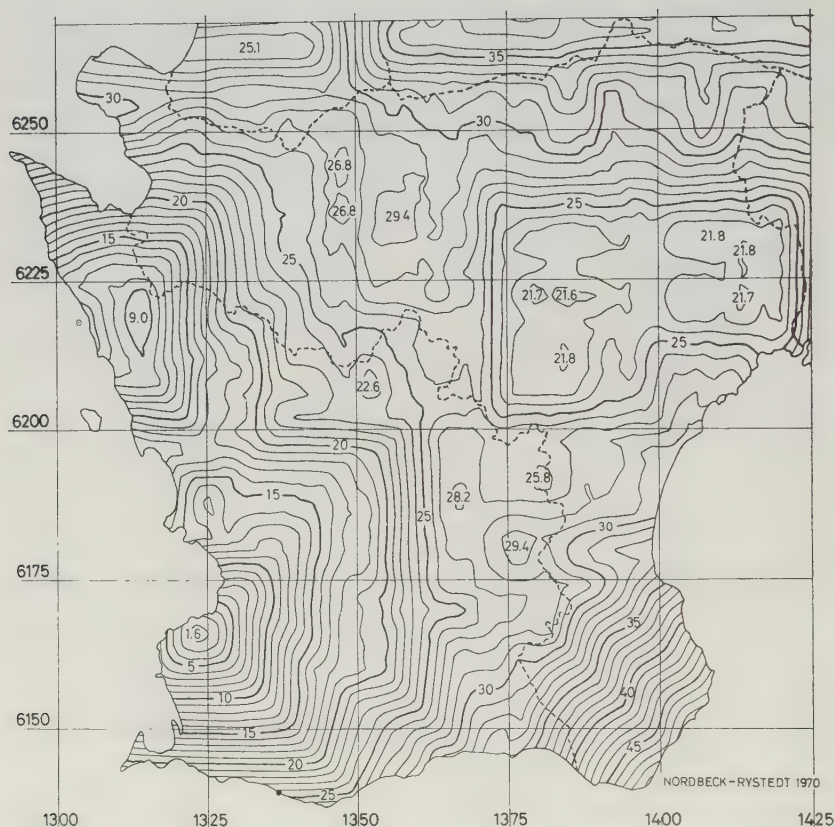


Figure 10.4. Range map of Skåne. Reference volume: 100 000 persons.

in the total population, the number of persons of school age in the total population or the number of pensioners also as a proportion of the total population. In the past, when preparing maps of this type data from local authorities and parishes has been used. This produced satisfactory maps due to the fact that the total populations of these areas were all of approximately the same size. The availability of coordinate based data has now led to its use in the preparation of fraction maps. When a numerical map of the NORI type is used as the basis for this operation highly unsatisfactory results are obtained. This is a result of the fact that population varies considerably from cell to cell (circle to circle) in this type of map. Thus the rule that approximately equal sized populations should be used when calculating fractions has been broken. Of the types of map described in this report the range map with a constant reference volume is the only type capable of providing a satisfactory basis for the construction of a fraction map.

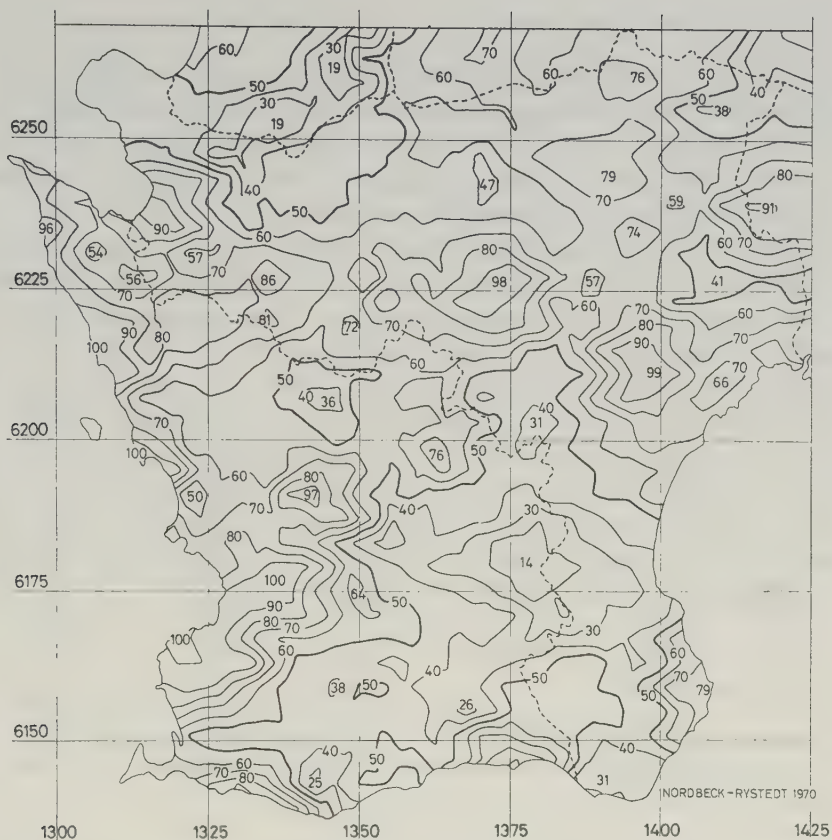


Figure 10.5. Fraction map of Skåne. The map shows the percentage of urban residents occurring amongst the 10 000 persons living nearest to every grid point. (Circular reference area, reference volume 10 000 persons). Equidistance: 4 km.

Figure 10.5 is a fraction map of Skåne showing the proportion of the population living in built-up areas in 1965. The reference volume is 10 000 persons. This map also shows the degree of urbanisation in different areas of the region. The western and southwestern areas are highly urbanised with up to 100 % of residents in built-up areas. In eastern Skåne the proportion of residents in rural areas is larger, giving a proportion of residents in urban areas as low as 14 %. A criticism of this particular map is that a reference volume containing 10 000 persons is far too small. A reference volume containing 100 000 persons has also been tried, the results from which exhibit the same tendencies as those of figure 10.5. The westerly and southwesterly parts of Skåne are highly urbanised whereas the easterly parts are only marginally urbanised.

10.6 Transport cost maps

Range maps are in themselves a useful aid when locating facilities. This is particularly true when the population is approximately evenly distributed over the surface. In other cases range maps should be supplemented by transport cost maps. The function mapped in this case, $f(X,Y)$, is the cost of transporting a certain quantity of goods from point (X,Y) to every one of those persons lying nearest to the point in question. This calculation is carried out at the same time as the range is calculated. The method is based on the assumption that the population in a ring around the point is evenly distributed. The next stage is the calculation of the mean distance between the ring's population and point (X,Y) . The simplest way of describing the ring is as the difference between two concentric circles, of which the largest has the radius $D + dD$ and the smallest the radius D . The mean distance is referred to as MDC and is calculated using formula 10.1 in which the constant k is equal to the reciprocal of the area of the ring. The set A contains only those points lying within the area of the ring.

$$\begin{aligned}
 MDC &= \iint_A k \sqrt{X^2 + Y^2} \, dX \, dY \\
 A &= \{X,Y; D \leq \sqrt{X^2 + Y^2} \leq D + dD\} \\
 k &= (\pi(2D + dD)dD)^{-1}
 \end{aligned} \tag{10.1}$$

Evaluation of integral 10.1 gives as a result formula 10.2. Insertion of $D = 0$ and $dD = 1$ in this formula gives $MDC = 2/3$, which is the value of the mean distance to the centre in a circle of radius 1.

$$MDC = 2(3 D^2 + 3 D dD + dD^2)/(6 D + 3 dD) \tag{10.2}$$

Formula 10.3 is produced by carrying out the division in formula 10.2. If D is considered to be a large number the last term in 10.3 becomes small, and thus an approximate value of MDC when D is larger than 1 is given by $MDC = D + 0.5 dD$.

$$MDC = D + 0.5 dD + dD^2/(12 D + 6 dD) \tag{10.3}$$

As has been pointed out earlier, square reference areas make possible the production of range maps in half the time necessary with reference circles. It is therefore desirable to carry out the calculation of mean distance for a "square ring", defined as the difference between two concentric squares. The length of half the side in the smaller of the two squares is referred to as D , in a similar way the length of half the side of the larger square is referred to as $D + dD$. The breadth of the ring is thus equal to dD . The notation is in principle the same as that used with reference circles. The mean distance from all points in the square ring to the centre point is referred to as MDK and is calculated using formula 10.4.

$$MDK = 1.1478 (D + 0.5 dD + dD^2/(12 D + 6 dD)) \tag{10.4}$$

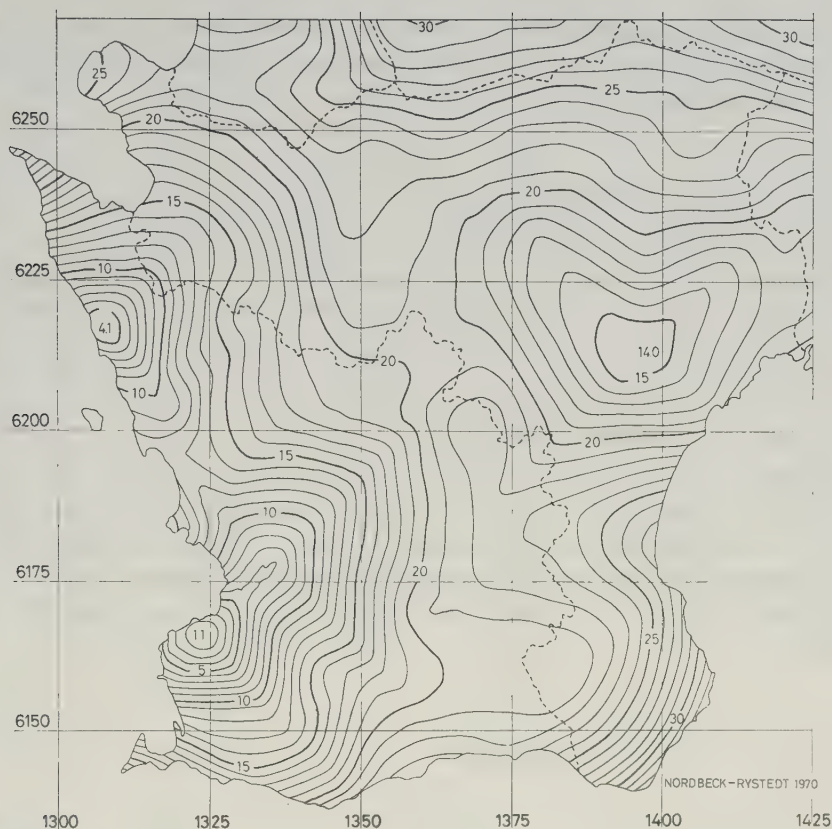


Figure 10.6. Transport cost map of Skåne. The map actually shows the total transportation distance from every point to the 100 000 persons living nearest the point in question. The reference area used in this case is not a circle but a square, the same method as was used for figure 10.4. Unit: 100 000 km.

When the mean distance has been calculated the next stage is to look up a table of transport costs in order to obtain the transport cost, *COST*, for the transportation of a certain quantity of goods the mean distance. The transportation cost can also be calculated from a transport cost function. The contribution from the population of the ring to the transport cost incurred by the population of the reference area, is given by multiplying the transport cost, as calculated, by the population of the ring.

The transportation cost, *COST*, is often expressed by the function shown at 10.5, where *A* is a fixed cost for loading and unloading, *D* is the distance the goods are to be transported and *B* is an exponent usually equal to or less than 1. If *B* is less than 1 one obtains a decreasing cost per kilometre the longer the distance travelled. *B* can also be seen as a shrinking factor

which shrinks longer distances more than shorter ones. C can thus be treated as a cost per kilometre, which becomes clear when B is given the value 1.

$$COST = A + C \cdot D^B \quad (10.5)$$

In the case $B = 1$, $COST$ is described by a linear function where C is the direct transport cost per unit distance. If transportation costs are being studied, in order to compare transportation costs at different points in the map, then the fixed loading and unloading cost A does not need to be included in the calculation. This is due to the fact that A is the same at all points (X,Y) . The cost per kilometre is also unimportant when studying linear transportation costs. Only the distance D is left. Thus one can just as well calculate the total transportation distance as the total transportation cost. Figure 10.6 is a transport cost map where the total distance has been substituted for the total transportation cost, as described above. The reference volume contains 100 000 persons. This map shows that the best siting from the transport cost point of view is in the west and southwest of Skåne whereas the southeast part of the province has the highest costs.

10.7 The NORREM program

The program used for the calculation of population range and fractions is called NORREM. It also calculates the cost of transporting a certain quantity of goods from point (X,Y) to all those persons living nearest to point (X,Y) , defined in terms of the range function. The use of this program results in three different maps; an ordinary range map, a fraction map and a transport cost map.

10.8 Isarithmic population maps and the location of facilities

Situations occur when an isarithmic population map of the type produced by NORI is a useful aid when discussing the location of facilities. Many different types of service facility require a pool of customers or clients the size depending on the type of facility. In other words, at least a certain number of persons must live within a certain distance of a point before it is possible to locate a facility at that point. The size of this minimum pool of customers is often referred to in location discussions as the threshold value for the service facility in question. Of course it is necessary to

realise that this threshold value is in fact a broad and diffuse interval. The distance referred to above is sometimes called the range of the facility in question, and is equal to the radius of the reference circle. The range of a facility is also difficult to state in precise terms.

If a choice of value for both threshold and range can be made then the NORI program can be used to map the area in which the facility is to be located. The resulting map shows the number of persons living within the given distance from a point (X,Y) . If, on the other hand, the question is how long is it necessary to go from point (X,Y) in order to ring in a certain number of persons, then the answer is given by using a population range map. Thus population range maps are also useable as a basis for location decisions.

11 Population potentials

The population potential function is quite abstract when compared with such functions as population density, population range and transportation cost. The following explanation will perhaps help to make it more concrete. One starts with a population distributed over an area. One can number every residence, $B_1, B_2, B_3, \dots, B_N$. In addition, one has a point, PI , which lies at distance, $D_1, D_2, D_3, \dots, D_N$, from each respective $B_1, B_2, B_3, \dots, B_N$. It is known that the probability that a person living at point B_J will visit point PI is dependent upon the distance D_J . The larger D_J becomes the smaller the probability. Stewart (1947) who first coined the term population potential discovered that this probability was inversely proportional to the distance D_J , as did Zipf (1949). A simple numerical example shows how this works. One has two points B_1 and B_2 each having a population of 1000. One also knows that the distance D_2 is 5 times as long as D_1 . If the probability that a person from B_1 visits point PI during a certain period of time is 0.25 the corresponding probability for point B_2 is 0.05. On the strength of the "law of large numbers" one can say that approximately 250 persons from point B_1 visit point PI during the time interval as do approximately 50 from point B_2 . The number of visitors to point PI from the points $B_3, \dots, B_N$ can also be calculated and the total number of visitors at point PI obtained by simple addition, for the time interval in question. The frequency of visits can then be calculated for all the points PI and the results presented in the form of an isarithmic map.

The point most frequently visited by the whole of the population can be determined by inspection of the isarithmic map. It is from a location point of view best to place a facility at that point if the number of visitors is to be maximized.

In most cases the precise number of visitors at the point in question is of no great interest. On the other hand the relationship between the frequency of visits at different points is of interest, i.e. that one point has twice as many visitors as another, which has in its turn three times as many visitors as another point, and so on. This makes it possible to avoid

calculating the exact probabilities as shown above. They are directly proportional to $1/D_j$ and this quotient can be used instead of the actual probability. This does not give the actual frequency but a function directly proportional to it. This new function is usually defined as the population potential and can be calculated using formula 11.1. P_i in this formula refers to the population potential at point PI . B_j is the number of persons residing at point BJ . D_j refers to the distance between points BJ and PI .

$$P_i = E_i + \sum_{j \neq i}^N \frac{B_j}{D_j} \quad (11.1)$$

It is evident from this formula that population potential is dependent on the units used to measure D_j . If kilometres are used one obtains a population potential 1000 times larger than that obtained when using metres. It is therefore necessary to give the units used for D_j when referring to population potential.

11.1 Weights other than distance

In formula 11.1 the populations B_j are weighted with the reciprocal of the distance, $1/D_j$, before the addition is carried out. This weight is assumed to be proportional to the probability $P(BJ:PI)$, that a person resident at point BJ will visit point PI . It appears highly unlikely that this probability is equal to V/D_j , where V is the coefficient of proportionality, that was assumed not to be necessary in formula 11.1. Several investigations exist which show that more accurate values of $P(BJ:PI)$ are obtained by using a power of D_j instead of the distance itself. In this case $P(BJ:PI)$ becomes V/D_j^B . The use of this value in the calculation of potential is shown in formula 11.2. If B is given the value 1 in formula 11.2 one obtains the same result as with formula 11.1. A program for the calculation of population potential would use formula 11.2 with the exponent B as one of the parameters.

$$P_i = E_i + \sum_{j \neq i}^N \frac{B_j}{D_j^B} \quad (11.2)$$

One can of course use other functions besides $P(BJ:PI) = V/D_j^B$ if it is so desired. In most cases the distance D_j is measured as straight line distance, but D_j can of course be equal to a distance via a transportation network, a travel time, a transportation cost and so on. Due to the fact that the different distances are usually simply related, the gain in the accuracy of the resulting potential is usually insignificant (Nordbeck & Rystedt 1969b).

Hence, the use of distance via a transportation network, instead of straight line distance has no effect other than a reduction of about 20—25 % in the resulting potential values.

The use of the total income of the settlement in question, instead of its population, can be used when calculating potentials (Warntz 1959). The result is what Warntz calls a gross economic population potential. Harris (1954) used the turnover in the retail trade when calculating the market potential throughout the USA. An analysis of Warntz' and Harris' maps shows that they are highly correlated with one another (Norborg 1962).

11.2 The contribution made by place PI to the population potential

Formulas 11.1 and 11.2 contain a term E_i which represents the contribution made by place PI , i.e. the place for which the potential is being calculated. The calculation of E_i involves the use of a number of simplifying assumptions. The first of these is that the area of the place PI is circular with a radius equal to R . The area of the circle is thus πR^2 . The assumption is also made, that the population B_i is evenly distributed throughout the area, thus making the density of population in area PI equal to $B_i/\pi R^2$. A ring of radius r having a width dr has an area, A , equal to $\pi(r+dr)^2 - \pi r^2$ which can be simplified to give $A = \pi(2rdr + dr^2)$. The number of persons B_r resident within the ring is thus $A \cdot B_i/\pi R^2$. On the basis of the definition of population potential given earlier the contribution P_r to the total potential P_i made by the population B_r is equal to B_r/r^B . Substitution, with the value for B_r obtained earlier gives formula 11.3. One then continues to construct new rings until the whole area of the circle of radius R has been covered.

$$P_r = \frac{\pi(2rdr + dr^2)B_i}{\pi R^2 r^B} \quad (11.3)$$

The contribution E_i made by place PI to the total potential is equal to the sum of the contributions P_r from the rings. In order to calculate E_i one lets dr go towards 0, integrating P_r . The result of this operation is shown by formula 11.4. Evaluation of the integral obtained gives formula 11.5. In this case the limitation is made that B should be smaller than 2.

$$E_i = \frac{2B_i}{R^2} \int_0^R r^{1-B} dr \quad (11.4)$$

$$E_i = \frac{2B_i}{R^2(2-B)} \left[r^{2-B} \right]_0^R \quad (11.5)$$

If B is equal to 2 then the integral $\int_0^R r^{1-B} dr$ in formula 11.4 becomes $\int_0^R r^{-1} dr$ i.e. E_i goes towards infinity when B goes towards 2 since the integral $\int_0^R r^{-1} dr$ is equal to $(\log R - \log 0)$. If B is allowed to be larger than 2 one obtains a division with 0 when substituting in the lower boundary. Thus even in this case E_i is undefined. On the other hand if B is smaller than 2 substitution in formula 11.5 gives formula 11.6 which can be used to calculate the contribution E_i made by the place PI .

$$E_i = 2B_i / ((2 - B)R^B) \quad (11.6)$$

If B is given the value 0 in formula 11.6 one discovers that E_i is equal to B_i , i.e. the contribution is equal to the actual population. Giving B the value 0 makes the calculation of potential independent of distance. This gives a population potential equal to the sum of the different populations included in the calculation.

Use of $B = 1$ in formula 11.6 also simplifies the calculation of E_i , as in this case it is only necessary to divide the population B_i with half the radius R . Occasionally it is stated that the average distance between persons living at the location in question is $R/2$ (Norborg 1962). This value is then used in the calculation of E_i . Unfortunately this is incorrect, which can easily be shown in the following way. The centre of a circle has, according to definition, the shortest mean distance to all other points in the circle. Formula 10.3 gives the value of the mean distance from the centre as $2R/3$ which is greater than $R/2$. If the mean distance from a point at the centre of the circle, to all the other points is equal to $2R/3$, then the mean distance for other points must be greater than this value. The mean distance between individuals is completely irrelevant in connection with the calculation of population potential.

The population at the place PI is a very important factor in population potential calculations (Norborg 1962). It is often so, that the contribution made by the population at the place PI is larger than the contribution made by all the other locations put together. The choice of the correct value for B is therefore of great importance. Claeson (1964) states that use of the simple function $P(B_j; PI) = V/D^{B_j}$, which forms the basis of formula 11.2 is unsatisfactory. It should be divided into at least two parts, with one B value for the short distances and another for the long distances. Thus the calculation of population potential could use one B value when calculating the contribution made by the place PI and another when calculating the other contributions.

11.3 Square net maps and the procedure NORPOT

In the preceding section it was assumed that population information was available for different administrative areas. It is, of course, possible to handle the situation where data is given in the form of a square net map, in exactly the same way. The advantage of a square net map is, that the data is organized in smaller units, thus resulting in more accurately calculated potentials. One is not in any way limited to a certain type of grid net, but as population potential is often presented in isarithmic map form a triangular grid with the base of the triangle equal to the height is highly suitable. Thus the result of the calculations is a numerical map of a type suitable for further processing with NORIP to give an isarithmic map.

When calculating the contribution made by the grid square in question to the potential the square is best treated as a circle of radius R . The square can then be handled in a similar manner to that described in the previous section. Development of a formula equivalent to that shown at 11.3, but applied to a square is necessary if this approximation is unacceptable. The use of such a formula would not appear to give any great advantage and thus complicates the calculations unnecessarily. In this connection it should be pointed out that a number of simplifying assumptions have been made prior to the calculation of population potential. These simplifications tend to have an influence on the resulting values greater than that resulting from the approximation of a square with a circle. Using the approximation of a circle with a square it is possible to obtain formula 11.7 from formula 11.6 by substituting for R with $S/\sqrt{\pi}$.

$$E_i = \frac{2B_i \pi^{0.5B}}{(2-B)S^B} \quad (11.7)$$

The contribution E_i made by the grid square in question can be calculated on the basis of this new formula. B_i refers to the population of the square while S gives its length of side. The squares are stored according to the NORI system without any overlapping.

12 Location calculations

A facility is to be located according to some optimizing criterion. In the case of a service facility it is often desirable to minimize the total transportation distance traveled by the clients in reaching the facility. As an alternative to road distance, one can use time or cost. In the case where a factory is to be located, the cost of transporting its raw materials, and the cost of distributing the finished products, also need to be minimized.

12.1 Location of one facility

In the first case, the simple situation where one facility is to be located in a position which minimizes the total distance travelled by its clients in reaching it is handled.

Transportation costs are assumed to be a linear function of the straight line distance between the customer and the facility. This reduces the problem to one of minimizing the customers' total travel distance. The frequency with which the facility is visited is assumed not to vary from customer to customer, i.e. a customer who lives near the facility does not visit it more often than a customer living further away. A square net map, such as figure 12.1, is used to describe the spatial distribution of the customers. The number of clients living in one square in the map is associated with the centre point of the square. The calculation is begun by assuming that the facility is to be placed at the bottom left hand corner of the first square of the first row in the map. The total cost incurred in transporting all the customers to that point is then calculated. This procedure is repeated using the second square in the row, and so on. The result of the calculations is a map showing how the function, $f(X,Y)$, describing the total transportation costs, varies over the surface. See figure 12.2. A map of this type shows immediately where the best location, from the point of view of minimizing transportation costs, lies. A further result is the amount by which transportation costs are raised, when locating the facility at a point other than the best.

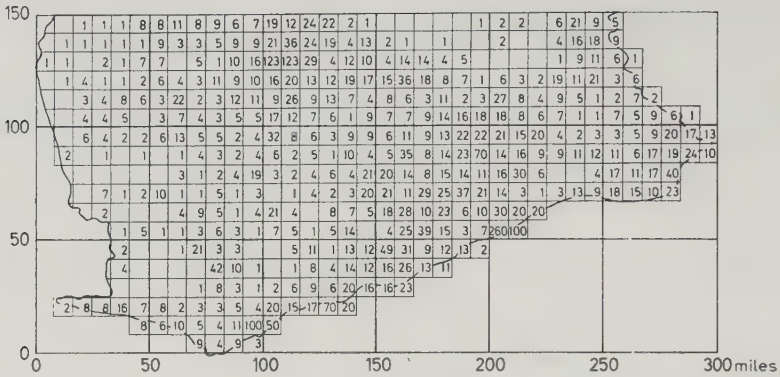


Figure 12.1. A square net map showing the distribution of population in southern Ghana. Grid size: 8.333 . . . miles. Source: Claeson, Godlund & Örtenblad 1969.

In the case of a factory, where both raw material transportation costs and delivery costs must be taken into account, the problem is only slightly more complicated. The transportation cost function used in this case contains two parts, one for raw material transportation and one for delivery.

12.2 The numerical relationship between the calculation of transportation distance and potentials

The technique used when calculating total transportation costs is very similar to that used when calculating population potentials. The NORPOT program calculates the potential P_i at a grid point G_i by summing the contributions made by all the populated grid squares. In the simplest case one divides the population of the square B_j with the distance D_j between the centre point of square j and grid point G_i , see formula 11.1. If one wishes to divide by a power of D_j i.e. D_j^B the exponent can be negative and multiplication carried out as shown in formula 12.1. If B is given the value -1 in this formula the formula is identical with that of 11.1. If B

$$P_i = E_i + \sum_{j \neq i} B_j \cdot D_j^B \quad (12.1)$$

is given the value $+1$ the corresponding total transportation distance is calculated. Thus, it is only necessary to change the value of B in the NORPOT procedure in order to calculate total transportation distances. Although the two functions are very similar, from a calculation point of view, there are clear differences between them. The total transportation

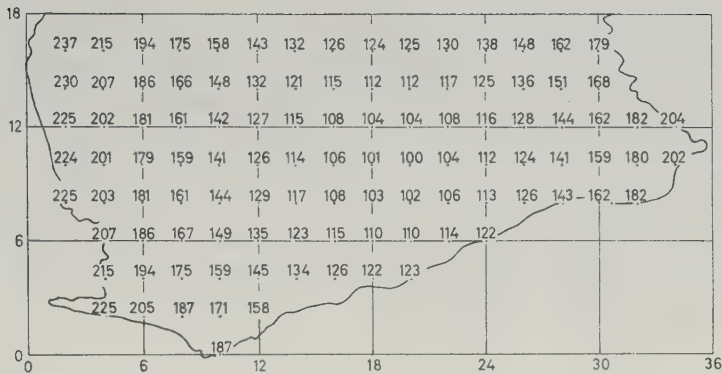


Figure 12.2. A numerical map showing the cost incurred when transporting the population of southern Ghana to different grid points. The cost is expressed as a percentage of the lowest cost occurring.

distance is a typical convex parabolic function. Thus, it has only one minimum point. The isarithms describing a function of this type are practically concentric rings. The potential function has a number of maximum points, also it is much more sensitive to the quality of the original data, i.e. it is highly dependent on the size of the grid net forming the basis for the calculation.

12.3 Location of one facility, iteration method

Unfortunately NORPOT is a very slow program, irrespective of whether it is used for location or potential calculations. The number of grid points handled can be referred to as n . The time required is then directly proportional to this n . Fortunately another, faster method of calculating the point where transportation distance is minimized exists. At the beginning a starting point is chosen, (X_s, Y_s) , usually a point lying near the centre of the area under study. The total transportation distance, $f(X_s, Y_s)$, is then calculated for this point. The next stage is to construct a line $Y = Y_s$ through the starting point parallel with the X axis. The calculation is then repeated at a number of points lying on this line. In this way one can determine a point, (X, Y_s) at which the total transportation distance is at a minimum. The calculations are then repeated, using this minimum point as a new starting point, with the difference that X is held constant and only Y varied. The whole procedure is repeated until the definitive minimum point is arrived at, having a value of the function less than at all the points surrounding it.

The position chosen for the starting point is particularly relevant to the time required to find the minimum point. If, due to luck, the point chosen is the same as the minimum point the calculation needs only be performed a small number of times. The choice of a point lying near the corner of the grid net makes it necessary to calculate the transportation distance function approximately $\sqrt{n}$ times before the result is obtained. In this case n is the total number of squares in the population map. The program using this iteration method is called NORLOC and this program is very much faster than the NORPOT program.

12.4 Location of several facilities, iteration method

The iteration method, described in the previous section, can be extended to carry out the location of two or more facilities, within the same area. In this case the aim is to find a point which minimizes $f(X_1, Y_1, X_2, Y_2, \dots, X_N, Y_N)$, where (X_1, Y_1) , (X_2, Y_2) , . . . , (X_N, Y_N) are the positions for N facilities.

The calculation of the best position for each of N facilities requires the choice of N starting points. The iteration is carried out in the following manner; the position of the first facility is varied in the manner illustrated in figure 12.3, whilst the remaining facilities are held fixed. In this way an optimum location for facility number 1 is obtained, for the situation where all remaining facilities are at fixed positions chosen in advance. The next stage is to make the new position obtained for facility number 1 into its new starting position and to proceed to vary the position of the second facility until an optimum location is obtained. The procedure is repeated for each of the facilities in order, holding all the other facilities fixed. In this way new positions are obtained for all the facilities, which are somewhat of an improvement on the original starting points.

When the first iteration is completed a second is begun, by once more moving facility number 1 in the manner shown in figure 12.3, repeating the operation for all the facilities in the same way as was done in the first iteration. The second iteration results in a further improvement to the positions occupied by the facilities. The iteration process is repeated until all the facilities are stable in their locations. A stable location is defined as that point a deviation from which causes an increase in the total transportation quantity occurring when all clients travel to their nearest facility.

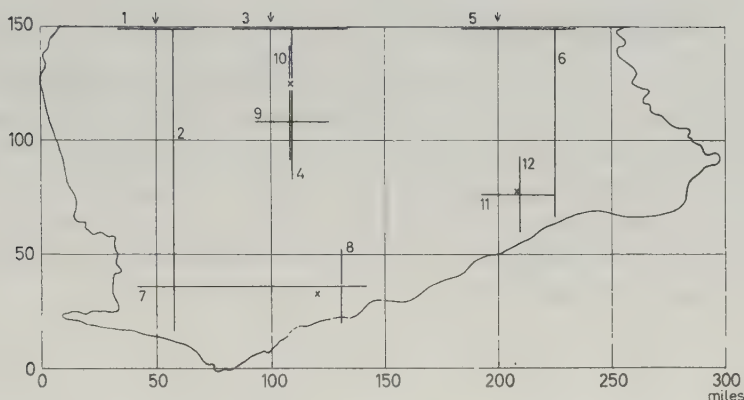


Figure 12.3. Approach used by the multiple location algorithm NORLOC. Every variable in the total transportation cost function is held fixed with the exception of one. This single variable is then varied until the value of the function is minimized. This variable is then held fixed and another varied. The procedure is repeated until a variation in the positions occupied by all the facilities does not lead to any reduction in the value assumed by the total transportation cost function. The arrows indicate the starting points chosen for three facilities, the crosses indicate the final positions they took up. The lines 1 to 6 show how the locations taken up by the facilities varied during the first iteration. In the same way, lines 7 to 12 belong to the second iteration.

12.5 The transportation cost function $COST$

Transportation costs are often given in a transportation cost table, but in most cases this table can be replaced by a function of type $A + CD^B$. In this expression A is a basic cost applied to all journeys irrespective of length. The variable D refers to the distance a client has to move in order to visit the facility lying nearest his home. The parameter B is an exponent, usually less than 1, due to the fact that transportation cost per kilometre usually decreases the longer the distance travelled.

A more complicated transportation cost function can of course be used. The actual distance via a road network can be used as an alternative to the straight line distance. The route via the road network can be chosen to be the fastest or the cheapest. This route can be obtained by application of one of the shortest route procedures included in NORMAP i.e. NORAN or NORCAS. This approach complicates the location of facilities program considerably. It is possible that the use of straight line distances does not affect the accuracy of the result obtained to any great degree, due to the different errors cancelling one another out. This should be investigated more closely, using actual data, in order to give a positive answer.

12.6 The NORLOC program

The program used to locate N facilities in the best locations from a transport cost point of view is called NORLOC. The program uses the previously described subroutines NORK, MAPOUT, and MAPIN. The subroutine NEWLOC calculates a total transportation cost when the facilities are located at given points. When the computer has decided on N optimum locations NEWLOC is called again, which results in the calculation of a map of the umlands.

13 The shortest distance between two nodes in a road network

In most cases the use of straight line distance gives a sufficiently accurate result, but situations do exist where a more accurate model is required. One method of increasing accuracy is to use actual travel distance instead of straight line distance. The term actual travel distance is, unfortunately, not as well defined as the term straight line distance. It would appear reasonable to define the actual travel distance between two nodes in a road network as the shortest distance between these two points via the network. The problem is thus one of finding this route, or alternatively its length.

There are a large number of methods which can be used to find solutions to shortest route problems. (Nordbeck & Rystedt 1969 b). These methods can be classified in two main groups; matrix methods and tree methods. The latter are sometimes called tree searching methods. Generally it can be stated that the matrix methods are easy to program but require the use of a great deal of core storage. These methods calculate the values for every pair of nodes in the network. A matrix method, the cascade method, is described below.

Tree methods are usually more difficult to program but require considerably less core storage capacity. It was possible to use a version of the tree method, the ellipse method, to process Lund's road network, which contains about 1000 nodes, on the SMIL computer which had a total core storage capacity of 8 K words. (Nordbeck & Rystedt 1969 b). The cascade method, applied to this road network, would have required a core storage capacity of about 1000 K words. The basic feature of the tree method is that the problem is handled one pair of nodes at time.

13.1 The cascade method and the NORCAS procedure

13.1.1 Road networks and matrices

The nodes in a network are numbered consecutively, without any special system being used. The road network can then be described in link form;

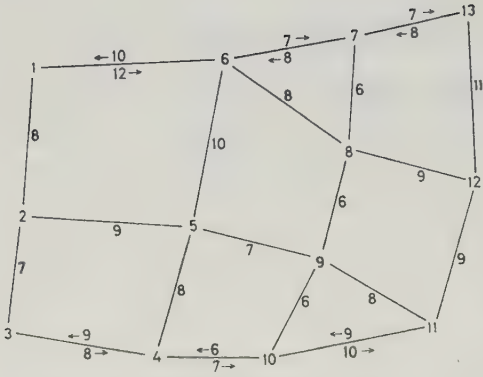


Figure 13.1. A simple road network with numbered nodes and the lengths of the links given.

every link being defined by the node numbers at its ends together with its length. Figure 13.1 shows a simple network including 13 nodes and 40 links. The network is not symmetrical, which means that there is at least one pair of nodes *AB*, where link *AB* is not as long as link *BA*. The network shown in the figure contains 6 such pairs of nodes. Table 13.1 shows the same network given in form of links.

A network can also be given in matrix form, see table 13.2. The values in the matrix describe the length of the respective links. A route between two nodes involving only one link is called a direct route, all other routes are referred to as indirect routes and are shown in table 13.2 by hyphens (-). A number of reasons indicate that it is best to represent the length of indirect journeys by a very large number. In the matrix shown as table 13.3 the hyphens are replaced by the infinity symbol (∞). This matrix describes a road network consisting of nodes 6, 7, 8, 12, and 13 in the network described in figure 13.1.

r	k	L	r	k	L	r	k	L	r	k	L
1	6	12	5	2	9	8	6	8	11	10	9
1	2	8	5	4	8	8	7	6	11	9	8
			5	9	7	8	12	9	11	12	9
2	1	8	5	6	10	8	9	6			
2	5	9							12	11	9
2	3	7	6	1	10	9	8	6	12	8	9
			6	7	7	9	11	8	12	13	11
3	2	7	6	8	8	9	10	6			
3	4	8	6	5	10	9	5	7	13	12	11
									13	7	8
4	3	9	7	6	8	10	9	6			
4	5	8	7	8	6	10	11	10			
4	10	7	7	13	7	10	4	6			

Table 13.1. The individual links in the road network shown in figure 13.1. A link is described by its end nodes (*r* and *k*) and its length (*L*). The road network contains 40 links, 14 pairs of which are symmetrical. Thus, e.g. link 2—5 is the same as link 5—2.

node r	node k												
	1	2	3	4	5	6	7	8	9	10	11	12	13
1	-	8	-	-	-	12	-	-	-	-	-	-	-
2	8	-	7	-	9	-	-	-	-	-	-	-	-
3	-	7	-	8	-	-	-	-	-	-	-	-	-
4	-	-	9	-	8	-	-	-	-	7	-	-	-
5	-	9	-	8	-	10	-	-	7	-	-	-	-
6	10	-	-	-	10	-	7	8	-	-	-	-	-
7	-	-	-	-	-	8	-	6	-	-	-	-	7
8	-	-	-	-	-	8	6	-	6	-	-	9	-
9	-	-	-	-	7	-	-	6	-	6	8	-	-
10	-	-	-	6	-	-	-	-	6	-	10	-	-
11	-	-	-	-	-	-	-	-	8	9	-	9	-
12	-	-	-	-	-	-	-	9	-	-	9	-	11
13	-	-	-	-	-	-	8	-	-	-	-	11	-

Table 13.2. The road network shown in figure 13.1 and described in table 13.1 given in matrix form. An element in the matrix is equal to the length of the corresponding link. In those cases where no direct connection exists between a pair of nodes in the network the matrix contains a hyphen (-).

13.1.2 The theoretical background to the cascade methods

A new matrix is constructed on the basis of the original matrix, in this case table 13.3. This is done by adding a number from the first row to a number in the first column, the resulting sum is entered in the matrix if it is less than the value already being there. In this case the first row and the first column remains unchanged. In the second row and the second column there is a very large number which is replaced by 15 which is the sum of 7 and 8. Other values in the second row remain unchanged. In the third row the third value is changed to 16, which is the sum of 8 and 8. The 4th and 5th rows remain unchanged. The resulting matrix is shown as table 13.4. In order to distinguish between this new matrix and that shown in table 13.3 the new matrix is referred to as the first stage matrix.

node r	node k				
	6	7	8	12	13
6	∞	7	8	∞	∞
7	8	∞	6	∞	7
8	8	6	∞	9	∞
12	∞	∞	9	∞	11
13	∞	8	∞	11	∞

Table 13.3. A road network consisting of nodes 6, 7, 8, 12 and 13 in figure 13.1. The network is given in matrix form with each element in the matrix giving the length of the respective link. In those cases where no direct connection exists between two nodes the matrix contains a very large number (∞).

node r	node k				
	6	7	8	12	13
6	∞	7	8	∞	∞
7	8	15	6	∞	7
8	8	6	16	9	∞
12	∞	∞	9	∞	11
13	∞	8	∞	11	∞

Table 13.4. The first stage matrix corresponding to the original matrix shown in table 13.3.

The first stage matrix is used to construct a second stage matrix. This is carried out in the same way as for the first matrix, with the difference that now the second row and the second column are used. The first value in the first row is very large and is replaced by 15, i.e. the sum of 7 and 8. The last value in the row becomes 14. The second row, like the second column, is unchanged. The third element in the third row becomes 12 instead of 16 and the last value in the row becomes 13. The 4th row remains the same as that in the first stage matrix. The 5th row is changed the most, the first value becomes 16, the third 14 and the 5th 15. The completed second stage matrix is shown in table 13.5.

A third stage matrix is then constructed on the basis of the second stage matrix. This is done by using the third row and the third column in the second stage matrix as the starting point for the calculations. The completed third stage matrix is shown in table 13.6. This matrix is the first which does not contain very large numbers. If one continues, with the 4th and 5th stage matrices, one discovers that they are identical with the third stage matrix and that the final matrix has been reached.

Study of the final matrix shows that it contains the length of the shortest route between every pair of nodes in the network. In this case the road network only contained 5 nodes, i.e. 6, 7, 8, 12, and 13 in the original network shown in figure 13.1. There is no problem involved in using much larger

node r	node k				
	6	7	8	12	13
6	15	7	8	∞	14
7	8	15	6	∞	7
8	8	6	12	9	13
12	∞	∞	9	∞	11
13	16	8	14	11	15

Table 13.5. Second stage matrix corresponding to the original matrix shown in table 13.3.

node r	node k				
	6	7	8	12	13
6	15	7	8	17	14
7	8	12	6	15	7
8	8	6	12	9	13
12	17	15	9	18	11
13	16	8	14	11	15

Table 13.6. Third stage matrix corresponding to the original matrix shown in table 13.3. This matrix is identical with the 4th and 5th stage matrices and is thus the final matrix.

matrices. The major result of the use of larger matrices is a greater number of intermediate steps and intermediate matrices. In the case of the matrix in table 13.1, 13 stages are required to obtain the final matrix shown in table 13.7.

An analysis of a final matrix shows that the distance from a node to the node itself is not equal to 0, which is surely the answer one expects. This is due to the fact that every route must contain at least one link, which makes a 0 distance impossible. This produces the result that the diagonal in the matrix shows the shortest distance from a node back to itself via another node in the network.

The technique for finding the length of the shortest route described above is usually referred to as the cascade method. The technique is simple to program and the resulting program is quick in operation. NORCAS uses this technique and, in those cases where only the length of the shortest

node r	node k												
	1	2	3	4	5	6	7	8	9	10	11	12	13
1	16	8	15	23	17	12	19	20	24	30	32	29	26
2	8	14	7	15	9	19	26	22	16	22	24	31	33
3	15	7	14	8	16	26	33	27	21	15	25	34	40
4	24	16	9	13	8	18	25	19	13	7	17	26	32
5	17	9	16	8	14	10	17	13	7	13	15	22	24
6	10	18	25	18	10	15	7	8	14	20	22	17	14
7	18	26	33	24	18	8	12	6	12	18	20	15	7
8	18	22	27	18	13	8	6	12	6	12	14	9	13
9	24	16	21	12	7	14	12	6	12	6	8	15	19
10	30	22	15	6	13	20	18	12	6	12	10	19	25
11	32	24	24	15	15	22	20	14	8	9	16	9	20
12	27	31	33	24	22	17	15	9	15	18	9	18	11
13	26	34	41	32	26	16	8	14	20	26	20	11	15

Table 13.7. Final matrix corresponding to the complete road network shown in figure 13.1. An element A_{rk} in the matrix is equal to the shortest distance between node r and node k .

route is required and the road network is fairly small, its use is recommended.

NORCAS uses the parameters N which refers to the number of nodes in the network and *SYMM* which describes whether the matrix is symmetrical or not. Giving *SYMM* the value 1 indicates that the matrix is symmetrical, a 0, on the other hand, indicates an unsymmetrical matrix. In the case of an unsymmetrical matrix the network is described to the computer in the manner shown in table 13.1, i.e. start node, second node and the length of the link between them. If the matrix is symmetrical the length in one direction is all that is required.

13.2 The ellipse method and the procedure NORAN

A tree method can also be used to determine the shortest route between a pair of nodes in a network. When using a tree method the two nodes are referred to as the start node, S and the goal node, G . The investigation is started by a search for route between S and G consisting of only one link. In most cases no such route is found. One then proceeds to search for a route between S and G consisting of two links and so on with 3, 4, 5, . . . links, until one has found a route linking S and G . If a node N can be reached in more than one way the shortest route, SN , is always chosen. A graphic presentation of all treated nodes and links reminds one of a tree or, more realistically, a bush and this is the reason why the tree method has received its name.

The method described in the previous paragraph gives a route SG containing the minimum number of links. The number of links is referred to as p and the route itself as SpG . There is of course no reason for supposing that route SpG is the shortest route between S and G . It is necessary to allow the tree to grow completely in all directions, until all routes from S are longer than the known route SpG , before the shortest route can be defined. One can be sure that SpG is the shortest route if no route between S and G shorter than SpG is discovered during the continued search. This can be stated in geometric terms as the investigation of all the routes from S to a node N which lies within a circle of radius equal to the length of route SpG , having S as its centre (Nordbeck & Rystedt 1969 b). See also figure 13.2. One can say that the tree method, as described, uses two restrictions; the tree restriction and the circle restriction. The circle restriction can in fact be replaced by a corresponding ellipse restriction. This is possible due to the fact that it is not necessary to search the routes between S and G via N where the sum of the two distances SN and NG is greater

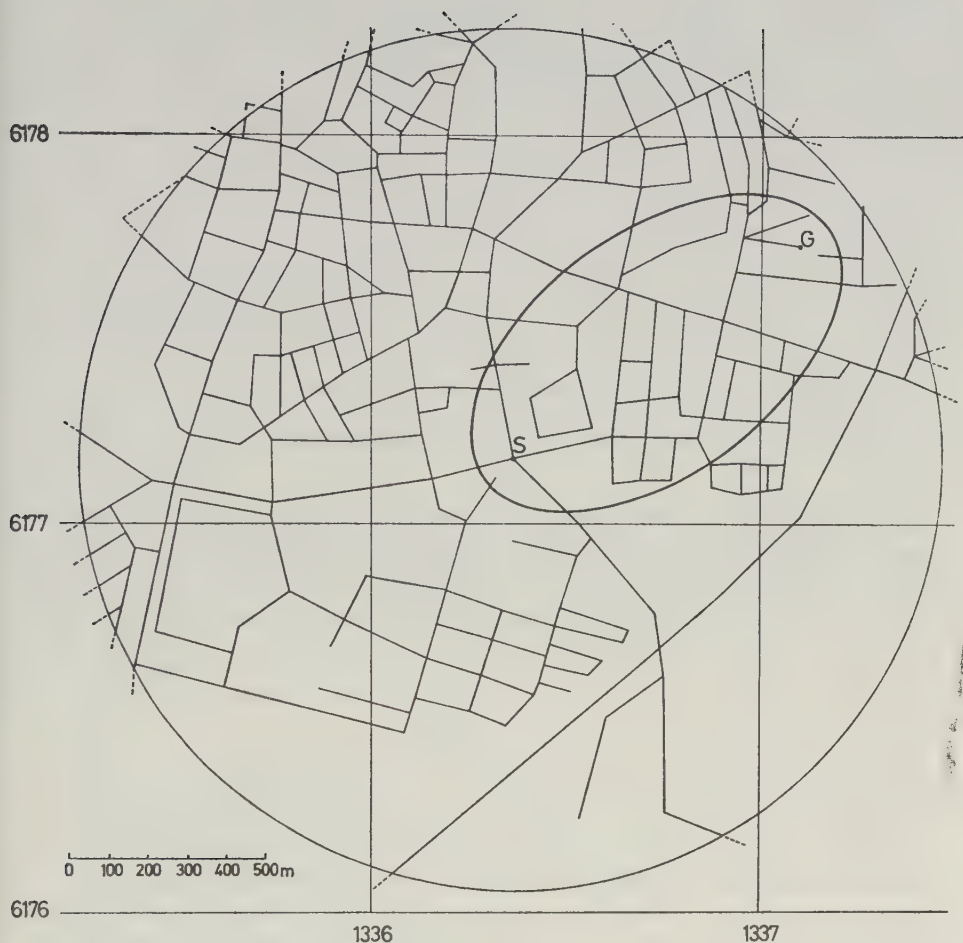


Figure 13.2. The shortest route circle having its centre at S and the corresponding shortest route ellipse with foci at S and G . The radius of the circle and the main axis of the ellipse are equal in length to the shortest route between points S and G via the road network.

than SpG . If one refers to the straight line distance, between S and N as SN and the corresponding distance between N and G as NG one obtains an ellipse with S and G as its foci. The length of the main axis of the ellipse is equal to the length of route SpG . Figure 13.2 shows that the number of routes requiring investigation is considerably reduced if the circle restriction is replaced by an ellipse restriction. A computer program based on the ellipse method is 15 to 20 times faster than a program using the shortest route circle. Tree methods and the ellipse methods especially are described in detail in two previous publications by Nordbeck (1964a)

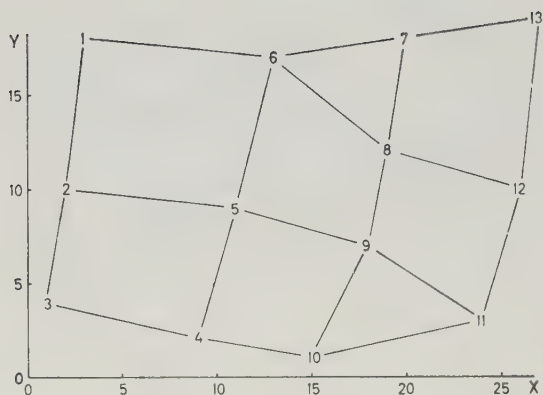


Figure 13.3. A simple road network placed in a coordinate system.

and Nordbeck & Rystedt (1969 b). It was not therefore considered necessary to describe them further in this publication. It should be pointed out, that use of the tree method results not only in the length of the shortest route, but also in the nodes it passes through. Thus the results can be used in traffic distribution models.

13.2.1 Coordinate based road networks

The ellipse method uses not only the actual length of the links but also the straight line distances, which is simply calculated using Pythagoras' theorem, with the prerequisite that the coordinates of the nodes in the road network are known. The assigning of coordinates to the nodes in a road network is carried out in precisely the same manner as the assigning of coordinates to the grid points in an irregular grid net. Thus, one begins by numbering the nodes consecutively, the coordinates are then measured and recorded for the node number in question, together with the node numbers of its direct neighbours. Figure 13.3 shows a road network placed in a 2-dimensional coordinate system. The same network is shown in table 13.8, but this time in the form of a coordinate based list. Such a list can of course be extended to include other information such as link length, road quality and road surface.

The restriction has been made that a node may not have more than 4 direct neighbours. If a node with more than 4 direct neighbours occurs in reality, it can be described as two nodes lying exceedingly close to one another, but having different numbers. This double numbering allows one to describe all junctions having 6 or less direct neighbours. Cases of more than 4 direct neighbours are exceedingly rare in reality. In Lund the road

i	X_i	Y_i	neighbours			
1	3	18	2	6		
2	2	10	1	3	5	
3	1	4	2	4		
4	9	2	3	5	10	
5	11	9	2	4	6	9
6	13	17	1	5	7	8
7	20	18	6	8	13	
8	19	12	6	7	9	12
9	18	7	5	8	10	11
10	15	1	4	9	11	
11	24	3	9	10	12	
12	26	10	8	11	13	
13	27	19	7	12		

Table 13.8. The road network shown in figure 13.3 given in the form of a coordinate based road register.

network has about 1000 nodes but only 3 of these have 5 or more direct neighbours.

13.2.2 Procedures included in the NORAN program

The program which uses the ellipse method on a coordinate based road network is called NORAN. It contains two subroutines LINK and MINRUT. LINK is used to determine which link $P1P2$ lies nearest to a point P . LINK also determines in which of the polygons created by the road network the point P lies. This means that LINK can also be used as a point-in-polygon program (Nordbeck & Rystedt 1969 b). The procedure also calculates the distance $D1$ between point P and point $P1$, which is one of the end points of the nearest link. A corresponding distance, $D2$ is calculated in the same way for the other end point $P2$. The description of the road network is then extended to include point P having points $P1$ and $P2$ as its nearest neighbours at distances $D1$ and $D2$ respectively.

The MINRUT procedure finds the shortest route between two nodes S and G in a road network.

The call on the NORAN procedure contains the coordinates of the starting point (SX,SY) and of the goal (GX,GY). Also included is a parameter MV which is equal to, or larger than, the radius of that circle which can be drawn around the largest polygon formed by the links in the road network. A further description of NORAN can be found in the publication referred to above. (Nordbeck & Rystedt 1969 b).

14 Barriers

The mapping and location programs described earlier in this report have, in the majority of cases, used straight line distance. The criticism can be made that this is too highly simplified and too far from reality. A more realistic model might be constructed using actual distance via a road network. In most cases, this is felt to be going further than the problem actually demands. The use of straight line distance between two points extended by multiplication with a suitable constant is satisfactory in the majority of cases. Problems arise when the two points lie on either side of a barrier of some kind. The use of straight line distance, even when extended, in cases where barriers exist produces a result far under the actual distance it is necessary to travel. Rivers, lakes, ranges of mountains, motorways and railways are examples of barriers. In all these cases the barrier is difficult to cross except at special crossing places, such as bridges, passes and viaducts. As an alternative to the use of the straight line distance between two points on either side of a barrier, the barrier itself, together with its crossing points, can be introduced into the problem. One could say that one forces the crow to take a route via a crossing point. Use of this method results in a number of straight line distances, one for each of the crossing points. Such a distance consists of two parts; the distance from the starting point to the crossing point and the distance from the crossing point to the goal. The shortest of the distances obtained is the shortest distance between the two points measured as the crow flies, but with the barrier taken into account. It is clear that this procedure considerably complicates the calculation of distances. Even so in certain cases it can be of value to include barriers when calculating distances. The next section describes an empirical example in which this was done.

14.1 Range maps, barriers and location calculations

Range maps can be a worthwhile aid when discussing the location of facilities (Nordbeck & Rystedt 1972), and their use was described in detail



Figure 14.1. A map of southern Ghana.

in chapter 10. Transportation cost maps are even more valuable for this purpose and the following empirical example describes their use. The presupposition made in the example was that a barrier of some kind existed, for example a river, also that this barrier was difficult or impossible to cross. The problem was then to determine the optimum location for a crossing point. In the case of a river the problem is to decide the optimum location for a bridge. The empirical data, used in the example, described an area in southern Ghana, where the Volta river and the Volta lake form a barrier. Figure 14.1 is a map of this area. Figures 14.2 A to D are a collection of transportation cost maps, constructed using the assumption that the Volta river is very difficult to cross. The reference volumes are 100 000, 250 000, 500 000 and 1 000 000 persons. Even the first of these maps indicates how bad the situation is in the eastern parts of the area. An enterprise requiring a threshold of 100 000 customers east of the Volta incurs transportation costs 10—20 times higher than a corresponding enterprise located in the Kumasi, Accra or Takoradi areas. This applies only to those enterprises not capable of exporting products eastwards to Togo.

The map shown in figure 14.2 B uses a reference volume of 250 000 persons. In this case the optimum location from the point of view of transportation costs is in the vicinity of Accra. Another local minimum can be seen in the vicinity of Kumasi. An enterprise in the eastern part of Ghana has, as one would expect, high transportation costs, although the relative increase is less than the increase occurring in the Accra area. Transportation costs were very low in the Accra area when a reference volume of 100 000 persons was used.

Further increases in the size of the reference volume cause the Volta barrier to become more and more apparent. This shows itself in the way the

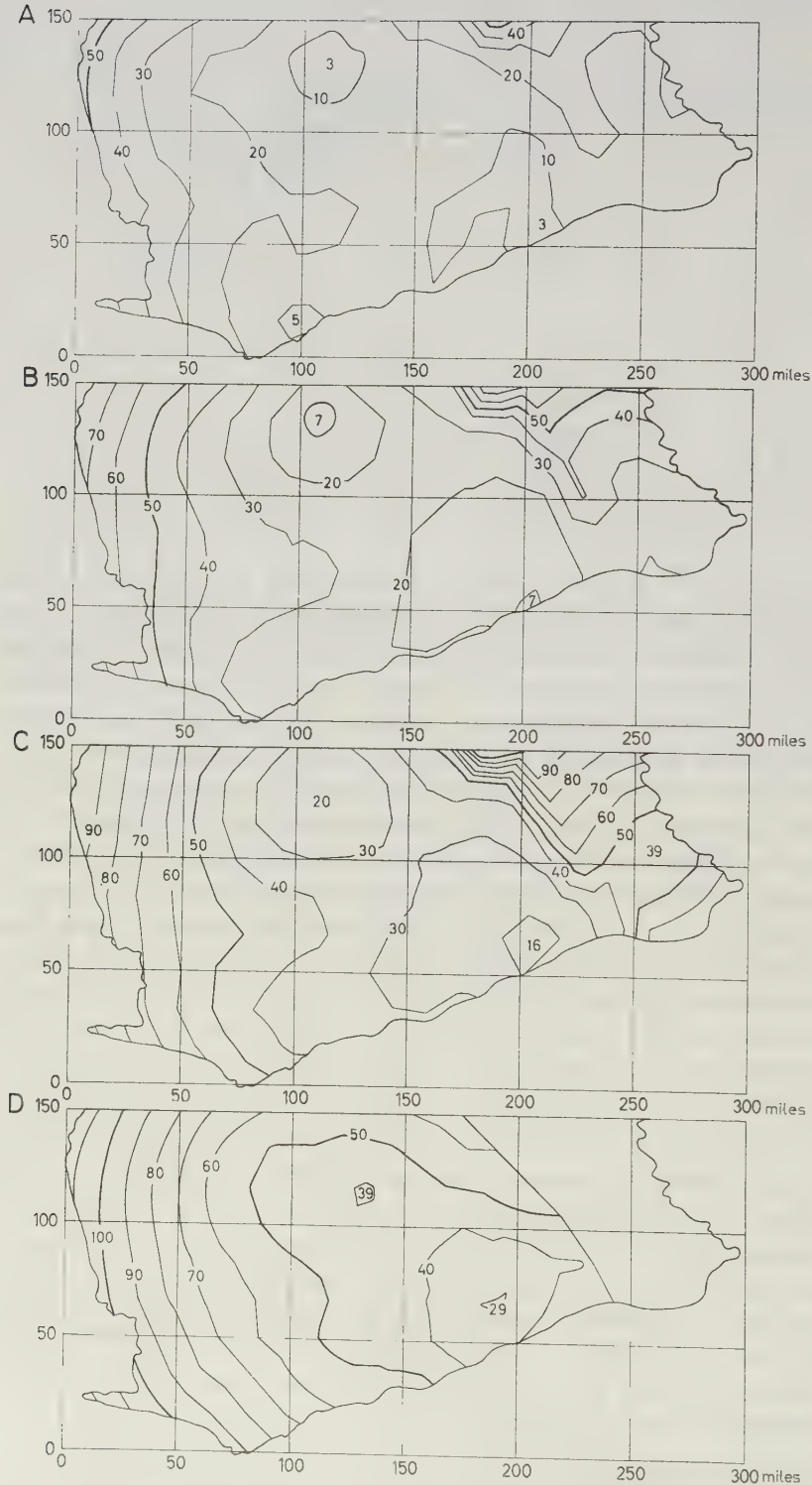


Figure 14.2. A collection of transportation cost maps of southern Ghana. The Volta river is assumed to be impossible to cross. Reference volumes: A: 100 000, B: 250 000, C: 500 000 and D: 1 000 000 persons.

isarithms orient themselves parallel to the river. Such a tendency appears quite clearly in figure 14.2 C and is even more evident in figure 14.2 D. In this latter case the reference volume is 1 000 000 persons. The population of eastern Ghana is less than this amount and as a result it is impossible to locate an enterprise there requiring a customer threshold of 1 000 000 or more persons.

Figures 14.3 A to D are a new collection of transportation cost maps, the reference volumes used are the same as before i.e. 100 000, 250 000, 500 000 and 1 000 000 persons. This time Volta is not impossible to cross, two bridges are available, one at Sogakofe and one at Akosombo. A comparison of figures 14.2 A and 14.3 A shows that the transportation cost picture is unaffected using 100 000 persons as the reference volume, with the exception of a small area in the central part of eastern Ghana, around Volta where a reduction in transportation costs by up to 50 % occurs.

When the reference volume is increased to 250 000 persons a clear improvement in the transportation cost picture occurs, especially in the northeastern part of the area, see figure 14.3 B. The tendency towards reduced transportation costs becomes stronger with increased reference volume. One can see from figure 14.3 C that an enterprise requiring a threshold of 500 000 persons, which is located in the eastern part of Ghana, acquires a considerably better market position when two bridges are available, as opposed to when no bridge is available. An enterprise requiring 1 000 000 customers can now be located east of the Volta river.

The final case is when the Volta is not considered to be a barrier. The river can be crossed without any problem at any point. Transportation costs sink in the area near the river. This shows itself in figure 14.4 as a displacement of the isarithms towards the east.

A series of range maps could also have been used to illustrate the variations (Nordbeck & Rystedt 1972). The transportation cost map is, of course, the better of the two, because the distribution of population within the reference area has been taken into account. A simple example illustrates this. Two points lie at the same range. In the case of the range map these two points are of equal value. In the case of one of the points the majority of the population are resident near to it and, as a result, the mean transportation distance is short. In the case of the other point the majority of the population are resident near the boundaries of the reference area and, as a result, the mean transportation distance is long. The difference between these two situations is considerable, even though the difference between their respective ranges is equal to 0.

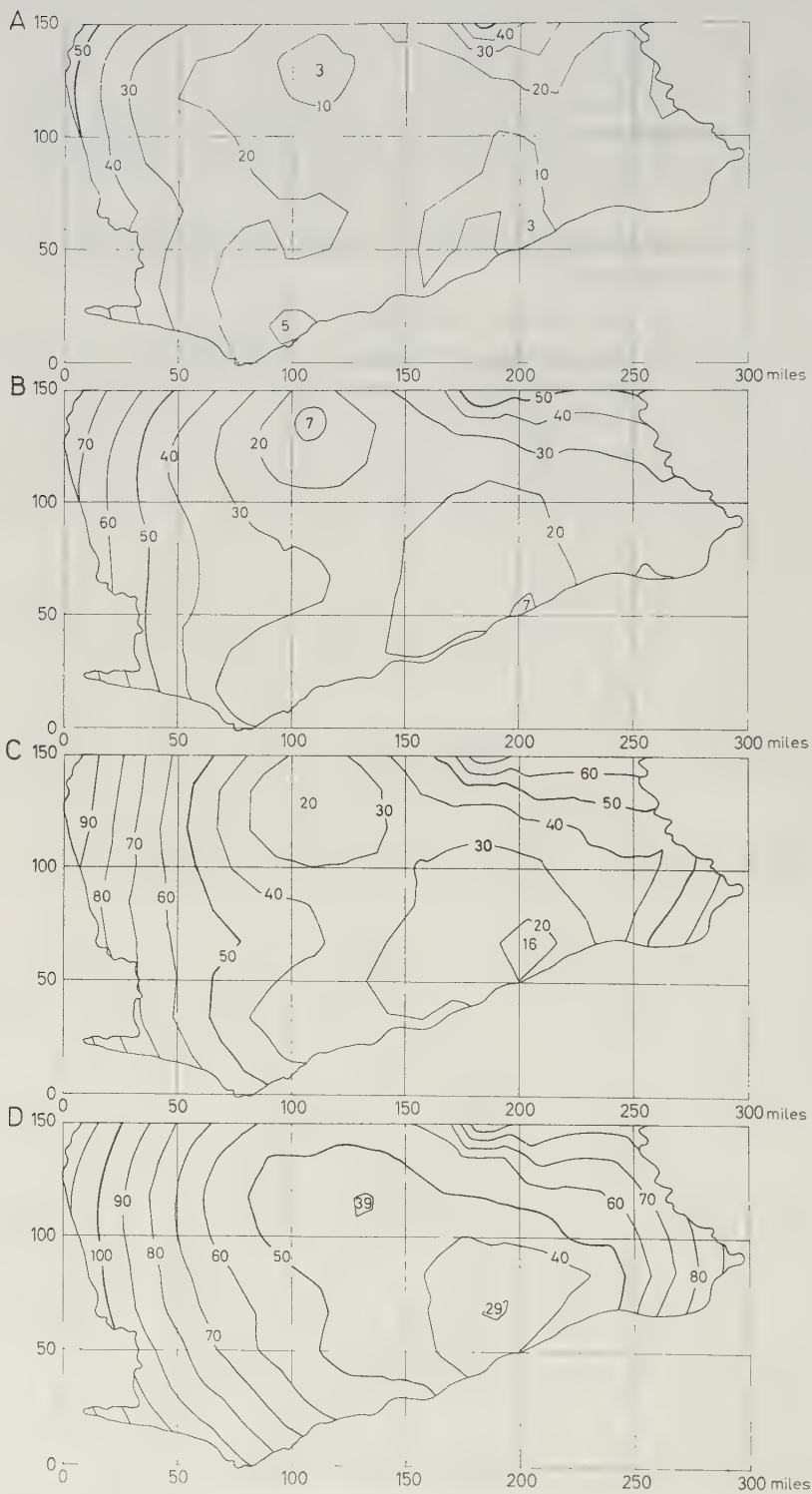


Figure 14.3. A collection of transportation cost maps of southern Ghana. The Volta can be crossed at Sogakofe and Akosombo. Reference volumes are the same as those of figure 14.2.

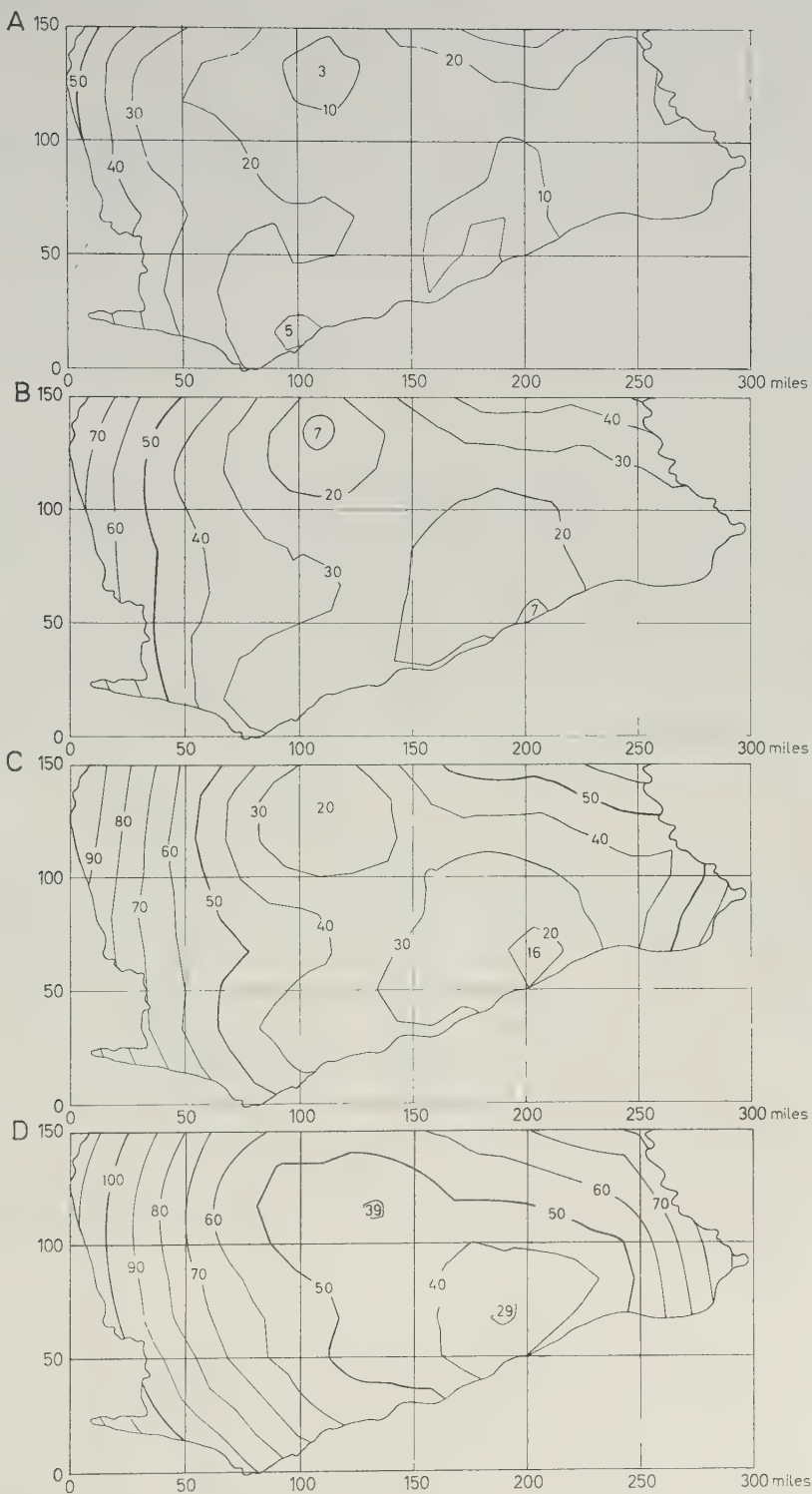


Figure 14.4. A collection of transportation cost maps of southern Ghana. The Volta is assumed not to present any barrier and can be crossed at any point. The reference volumes used are the same as those of figures 14.2 and 14.3.

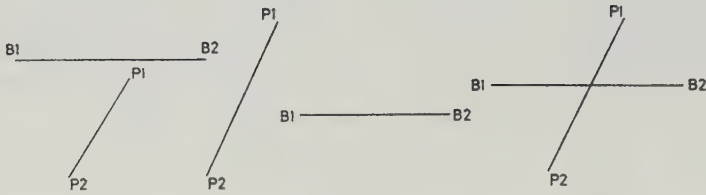


Figure 14.5. A line P_1P_2 is barred by line B_1B_2 only when points P_1 and P_2 lie on opposite sides of line B_1B_2 at the same time as points B_1 and B_2 lie on opposite sides of line P_1P_2 .

In the case of the bridge over the Volta the situation mapped was the situation as it exists, with the simplifying assumption that one could travel in all directions equally easily, with the exception of the Volta crossing. The two bridges exist in reality, but they could just as easily have been imaginary. Thus, it is possible to apply the technique when deciding the location of a new bridge. This is done by applying the range map program, using a number of imaginary bridge locations, and then choosing the location which gives the largest gain seen from society's point of view.

14.2 Straight line barriers and the NORLAS program

The simplest form of barrier is a straight line, defined by its end points, B_1 and B_2 . Figure 14.5 shows that a barrier between two points P_1 and P_2 exists only when P_1 and P_2 lie on opposite sides of the line B_1B_2 and points B_1 and B_2 lie on opposite sides of line P_1P_2 . In order to determine whether a barrier exists between points P_1 and P_2 it is necessary to begin by constructing triangles $B_1B_2P_1B_1$ and $B_1B_2P_2B_1$ as shown in figure 14.6. One then studies the orientation of these two triangles by calculating their area using the determinant formula. If the area of the two triangles has the same sign they are similarly orientated and no barrier exists. If, on the other hand, the two areas have different signs one proceeds to apply the determinant formula to triangles $P_1P_2B_1P_1$

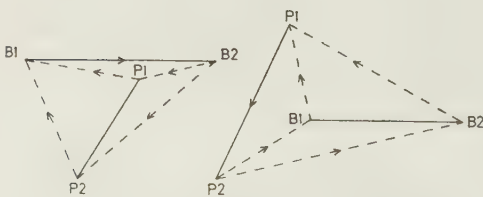
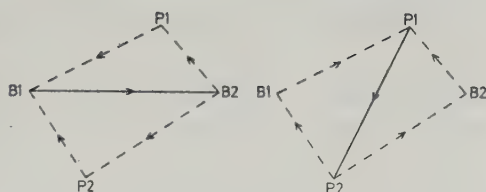


Figure 14.6. Two triangles are constructed using B_1B_2 as the base and points P_1 and P_2 respectively as the apices. If both of these triangles have the same orientation the line B_1B_2 does not form a barrier between points P_1 and P_2 .

Figure 14.7. In order that the line B_1B_2 should bar line P_1P_2 the triangles $B_1B_2P_1B_1$ and $B_1B_2P_2B_1$ should have different orientation as should triangles $P_1P_2B_1P_1$ and $P_1P_2B_2P_1$.



and $P_1P_2B_2P_1$. Only in the case where the orientation of these two triangles is different does a barrier exist. This case is illustrated in figure 14.7. When it has been determined that a barrier exists between points P_1 and P_2 one proceeds to calculate the length of the two routes $P_1B_1P_2$ and $P_1B_2P_2$. The shortest of these two routes is then chosen and used in all further calculations instead of the straight line distance P_1P_2 .

When the barrier is made up of a number of line segments with a number of crossing points $B_1, B_2, B_3, \dots, B_N$ every segment is individually handled in the way described above. As the result one obtains N distances, $P_1B_1P_2, P_1B_2P_2, P_1B_3P_2, \dots, P_1B_NP_2$, the shortest of which is then chosen and used in all further calculations.

The method of handling barriers described above can even be used when the barrier is given in the form of a simple polygon train. If the polygon train only contains two sides no extra problems are presented. In other cases the polygon is required to be uncomplicated and should only to a small degree deviate from a straight line in form. In the Volta example, described in section 14.2, the barrier was approximated with a polygon train containing three sides giving two crossing points.

It would appear possible to use other polygon trains than the very simplest as barriers and to combine several barriers in one problem. Unfortunately this is easier said than done. The problem becomes an exceedingly difficult combinatorial problem, for which no general solution appears to be available. In such cases it is simpler to work with a road network and a shortest route program.

The NORLAS program is used when the barrier is given as a straight line. The two points P_1 and P_2 have the coordinates (X_1, Y_1) and (X_2, Y_2) respectively. These coordinates are contained in the procedure call, in the above named order, together with a field S containing the 4 coordinates of the end points, B_1 and B_2 . The final parameter D gives the length of the shortest of the two routes $P_1B_1P_2$ and $P_1B_2P_2$.

14.3 Polygon barriers and the program NORPAS

It is desired to investigate whether a barrier in the form of a convex polygon $B_1B_2 \dots B_N$ lies across the straight line between two points P_1 and P_2 . In convex polygons all the internal angles are less than 180° . In cases where a barrier exists the straight line P_1P_2 crosses two sides of the polygon. Thus NORLAS can be used to see whether a polygon barrier blocks a line or not.

In those cases where a line is blocked by a polygon barrier it is necessary to go round it. To find the route it is necessary to take one begins by drawing straight lines between P_1 and all the corners B_V , ($V=1,2, \dots, N$), of the polygon. The angle made by P_1B_V with the line P_1P_2 is referred to as FIV , in this case V is an index. The corners of the polygon can be divided into two groups, on the basis of whether they lie to the left or to the right of line P_1P_2 . In a similar way two groups of angles can be obtained, the left hand and the right hand angles. The largest left hand and the largest right hand angle are then found by calculating the cosine of the angles using a formula given by Nordbeck (1964b). In the case shown in figure 14.8 one obtains angles $F12$ and $F17$. Thus, it is necessary to go from point P_1 to B_2 or B_7 in order to go round the polygon. The calculation is then repeated using point P_2 as the starting point. In this case one obtains angle $A3$ as the largest left hand angle and angle $A5$ as the largest right hand angle. Starting from point P_2 one goes either to corner B_3 or B_5 when going around the polygon. If one can proceed to the same corner of the polygon when starting from point P_1 as when starting from point P_2 the problem is solved, otherwise it is necessary to follow the sides of the polygon until one reaches the corner chosen from the opposite direction. In the example shown in figure 14.8 the problem is now to decide which of the two routes, $P_1B_2B_3P_2$ or $P_2B_5B_6B_7P_1$ is the shortest.

The NORPAS program is used when the barrier is given in the form of a polygon $B_1B_2 \dots B_NB_1$. The parameters used are, in principle, the same as those of the NORLAS program. If the barrier lies across the straight line route the value of variable D is calculated. D is the length of the shortest route around the barrier from point P_1 to point P_2 .

14.4 Combinations of several barriers and NORAS

It has already been pointed out that if several barriers are combined in the same execution of the program a difficult combinatorial problem presents

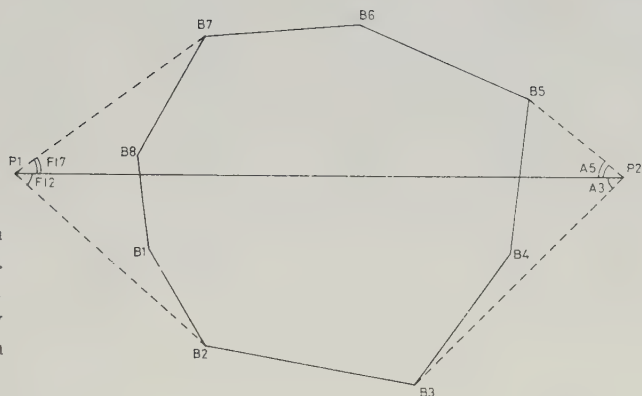


Figure 14.8. A polygon barring a straight line. The straight line $P2P1$ must be replaced by the polygon train $P2B5B6B7P1$.

itself. This problem is sufficiently complicated to make the development of a general solution nearly impossible. It is not sufficient to apply NORLAS or NORPAS a number of times to a series of barriers. Instead, the problem must be attacked from an entirely different direction. One approach is to construct an imaginary road network around the barriers, which is then searched for the shortest route using the MINRUT program. A simple explanation of this is given by the following example.

Figure 14.9 shows three polygons numbered 1, 3 and 4, together with a polygon train numbered 2. One can say that the barriers are built up of three closed polygons and one open. The closed polygons should be convex. A concave polygon is automatically replaced by a convex by the addition of triangles to the polygon. If it is absolutely necessary to use a concave polygon it can be built up from two or more convex overlapping polygons. In the case of the closed polygons the only way to pass the barrier is by going round it since the corners as well as the sides are impassable. In the case of the polygon train the sides are impassable but the corners are not.

The imaginary road network is built up by linking every corner in the different polygons with every other. The restriction is made that a linkage cannot be created if a barrier lies in the way. Figure 14.10 shows the road network obtained using the polygons shown in figure 14.9. The figure clearly shows that the result is highly complicated. A node in this network has an unlimited number of direct neighbours.

The imaginary network is then extended with two nodes, i.e. the start and finish points $P1$ and $P2$. Figure 14.11 contains these two points at the coordinate positions (14,3) and (10,28) and a straight line has been drawn

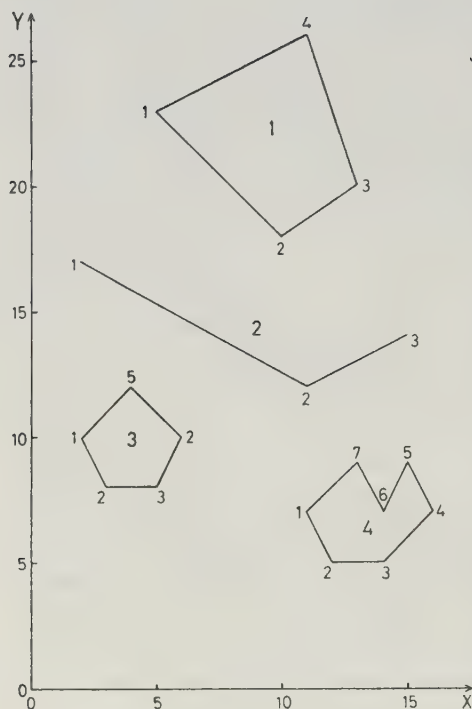


Figure 14.9. There exist 4 barriers within the area, three closed polygons, (1, 3 and 4) together with one open polygon, (2).

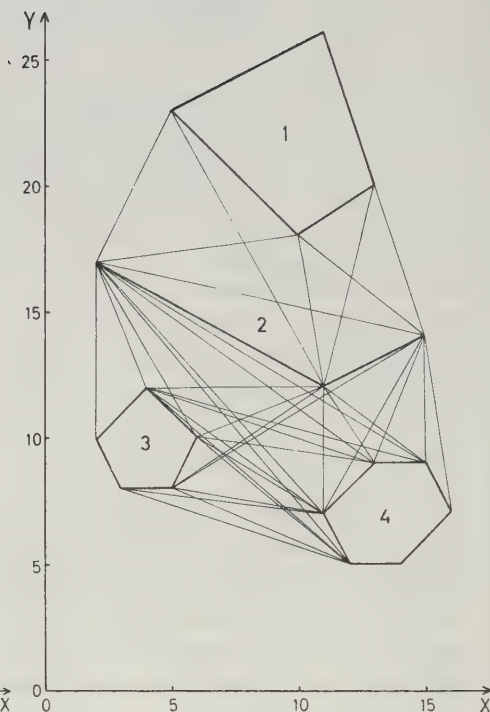


Figure 14.10. The imaginary road network obtained using the polygons shown in figure 14.9.

between them. This route is blocked by polygons number 1, 2 and 4. The road network is extended by constructing and entering into the road network all those lines between the two points and the corners of the polygons which are not blocked by a barrier. The imaginary road network used in the example consists of all the non-dotted lines in figure 14.10 and 14.11. Only now is it possible to find the shortest route between the two points which is not blocked by a barrier. This route is found by application of one of the shortest route programs MINRUT or NORCAS. In the example shown in figure 14.11 the shortest route goes via the polygon corners 4: 4, 2: 3, 1: 3 and 1: 4. This example illustrates clearly how complicated the use of barriers can become. A careful consideration should therefore be made before this technique is used. It is not at all certain that the advantages outweigh the disadvantages incurred when using the more complicated model. In most cases the inaccuracies which occur when no barriers are used tend to cancel one another out. In the case when a decision is to

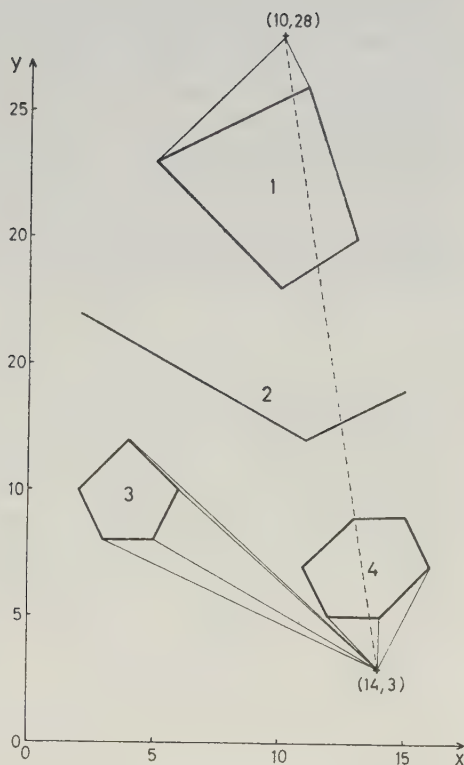


Figure 14.11. Those lines linking the start and finish points to the corners of the polygons which are not blocked by a barrier are entered into the imaginary road network shown in figure 14.10.

be made between the three different approaches to the measurement of distance one should first choose to measure straight line distance, second to measure via the shortest route through a road network and third, and last, to measure distance around barriers.

The NORAS procedure is used when several barriers built up of polygons are present. Due to the fact that it is an exceedingly long program it is only available in an ALGOL version. The program was felt to have too little application and to be too expensive in use to merit its translation into FORTRAN.

NORAS contains several versions of the procedures described above. One example is SPARR which is in principal the same as NORLAS. Another of these procedures is MINROT which differs from the MINRUT described previously in that it itself contains 4 subroutines X, Y, NONGBS, and INDEX. These procedures obtain data describing the node in question, X obtains the X coordinate, Y the Y coordinate, NONGBS the number of

direct neighbours to the node in question and INDEX the numbers of these neighbours and the length of the corresponding links.

The STORE procedure creates and stores the imaginary road network. NEWNP is used to include the start and goal points in the road network. CONVEX discovers whether the polygons forming the barriers are convex or not. If the polygons are concave they require to be extended using triangles to make them convex. NORPCO discovers whether a point lies outside, inside, or on the boundary of a convex polygon.

15 Sorting on the basis of polygons

Coordinate based data can be used for other purposes than map construction, for example the extraction of data relevant to a particular chosen polygon. In this way one can determine the population of a particular area, the advantage being that the population need not be that of a certain administrative district or combination of administrative districts. This is done by using a suitable point-in-polygon program to determine whether the coordinate based data point lies inside the chosen polygon. If such is the case the data is added to that already stored for the polygon. If population data is being used then one can obtain the total population of the polygon, also if one so wishes the data can be presented in different categories, such as age or sex. Nordbeck & Rystedt (1967) have described in detail a number of approaches to the point-in-polygon problem. As a result they will not be handled in great detail in this publication.

15.1 Tests for convexity in polygons

A convex polygon is defined as a polygon in which all of the internal angles are less than 180° . A concave polygon thus contains at least one internal angle greater than 180° . It is often necessary to determine whether a polygon is concave or convex. For example, a convex polygon can be processed using a considerably more simple point-in-polygon program than is necessary with a concave polygon. In the case of a convex polygon the triangles, which can be constructed using three succeeding vertices in the polygon all have the same orientation as the polygon itself. The orientation of a polygon can be found by calculating its area using the determinant formula.

The procedure which investigates whether a polygon is convex or concave is a logical FORTRAN function called CONVEX. The function acquires the value *true* when the polygon is convex and *false* when it is concave.

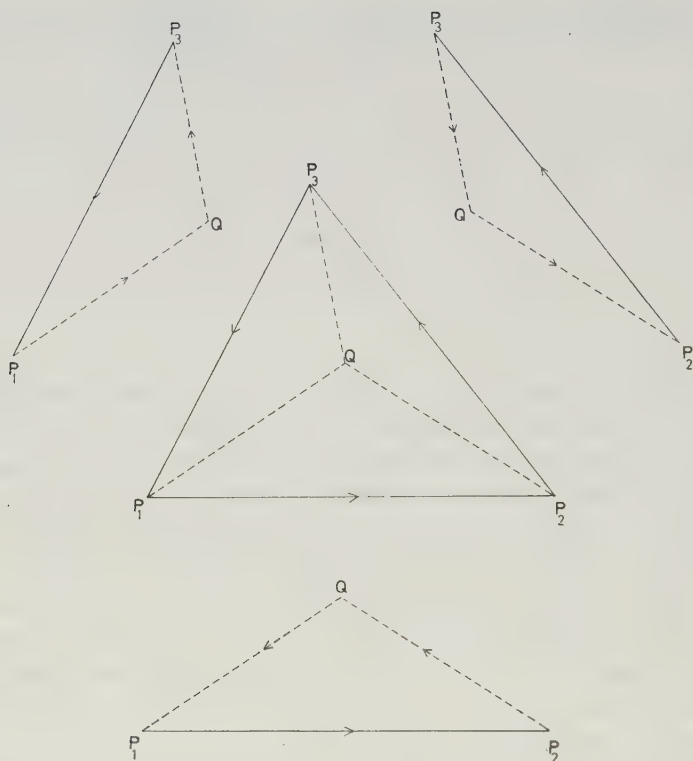


Figure 15.1. A point Q lies within a convex polygon $P_1P_2 \dots P_nP_1$, when all the triangles, $P_kP_{k+1}QP_k$ have the same orientation as the polygon.

15.2 Convex polygons and NORPCO

Several mathematical theorems exist, which can be used to determine whether a point lies inside or outside a polygon (Nordbeck & Rystedt 1967). In the case of concave polygons the most effective theorem is probably the orientation theorem. It states that a point Q does not lie within a polygon $P_1P_2 \dots P_n$ if one of the triangles $P_kP_{k+1}QP_k$ does not have the same orientation as the polygon. See figure 15.1. The point lies within the polygon only when all the triangles and the polygon have the same orientation.

The FORTRAN function NORPCO acquires the value (-1) when a point Q lies outside the polygon, $(+1)$ when the point lies inside and (0) when the point lies on the boundary of a convex polygon.

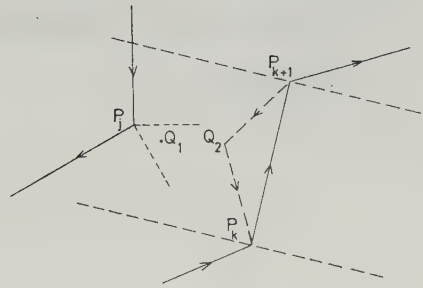


Figure 15.2. Illustration to the orientation theorem for concave polygons.

15.3 The orientation theorem for concave polygons and NORP

The orientation theorem, described in the previous section, only applies to convex polygons. A simple modification allows it to be applied to all polygons (Nordbeck & Rystedt 1967). As can be seen from figure 15.2 this version consists of two parts.

1. A point Q lies within a polygon P if it lies nearer the nearest vertex than the nearest side, in the case when the vertex is concave.
2. A point Q lies within a polygon P if it lies nearer the nearest side $P_k P_{k+1}$ than the nearest vertex, when the orientation of the triangle $P_k P_{k+1} Q P_k$ is the same as the polygon itself.

The extended orientation theorem applies to all polygons but it is very uneconomical when applied to convex polygons.

The FORTRAN function using the extended orientation theorem is called NORP. It uses the same parameters and input data and it acquires the same values as NORPCO.

16 Directional information and fabric diagrams

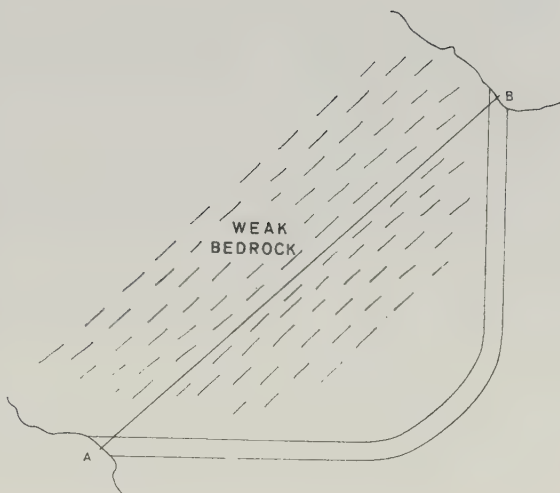
The programs and methods described here have been directed mainly towards community planning. Such problems as the use of coordinate based data in the mapping of population distribution, the distribution of traffic, or the choice of the most suitable location for a given activity etc. However, these methods have a much wider field of application than that provided by the mapping of population distributions. An example of a different type of application, using overlapping reference areas and isarithms is provided in this section. The example is a program for the automatic analysis of data describing the directions assumed by planes and lines. Planes of this type occur when describing tectonic faults.

A well known quotation states that one should build one's house upon rock. Associated with this quotation is the idea that rocks are permanent and fixed, to be moved only by a catastrophe. Unfortunately this is not the case. Rock often contains faults and other weaknesses, which can lead to displacements if the load is transferred in an incorrect manner. A quantitative and qualitative analysis of tectonic elements is desirable before a decision is made regarding the siting of buildings. Strengthening of the bedrock can be necessary, before foundation work can be carried out. In such a case a different siting for the building could provide a more economic solution. Such a case is illustrated in figure 16.1.

A tunnel is to be built between points A and B. Unfortunately, it has been discovered by a tectonic investigation that faults in the bedrock run parallel with the straight line AB. If a tunnel is built on this line the roof requires heavy reinforcement to avoid it caving in. If the tunnel is built in a zigzag line the extra reinforcement is not necessary. The overall result is a somewhat longer tunnel that, even still, can be cheaper than the straight tunnel between A and B.

Faults in the bedrock have a drainage function, too. If the direction of the faults is known the direction of drainage is also known. Figure 16.2 shows a case where a building is to be built partly excavated into the

Figure 16.1. A tunnel is to be constructed between points A and B. Faults in the bedrock run parallel with the straight line AB. Also the rock is weak in the vicinity of this line. It is therefore cheaper to give the tunnel another route than that along the direct line between the two points.



bedrock. The figure also shows the direction taken by the main drainage. One should, of course, locate the main drainage where it crosses the natural drainage lines in the bedrock. A tectonic analysis can in such cases lead to economies in the construction of the foundation.

One way of analysing directional information is to enter it into a fabric diagram. The construction of a fabric diagram by manual means is complicated, requires much time and is very boring, as a result it is expensive. A fabric diagram is in fact a sort of isarithmic map. It is thus possible to automate its construction, using the same principles as were used in the construction of isarithmic maps of population.

16.1 The description of the direction of a plane in space

The direction taken by a plane in space can be described with the help of two angles. Strike is defined as the angle between the north-south line and the line of intersection between the plane and the horizontal plane. See figure 16.3. If the plane strikes from the north-west to the south-east (or vice versa) the angle of strike is given in the form North-angle-West. In figure 16.3 A the plane strikes at the angle $N60^{\circ}W$. If, on the other hand, the plane strikes from north-east to south-west (or vice versa) the angle is given in the form North-angle-East. In figure 16.3 B the angle of strike is $N30^{\circ}E$. If strike towards the west is defined as positive and strike towards the east as negative the possible values fall between -90° and

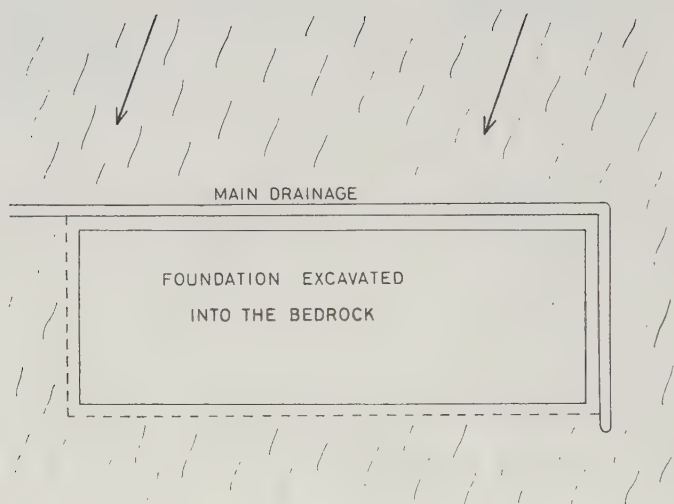


Figure 16.2. When the foundations of a building are excavated into the bedrock, they should be located across the fault and not parallel with them.

$+90^\circ$. As an alternative to the giving of angles of strike in the form N-angle-W or N-angle-E they can be given in numeric form using the + and - signs. The angle in figure 16.3 A is thus 60° , while the angle in figure 16.3 B is -30° .

Dip is defined as the angle between the horizontal plane and the plane in question. It is usually given using the points of the compass E, W, N, S, NE, SE, NW, and SW. The majority of these are actually unnecessary. E and W are usually sufficient, as all planes dip somewhat to east or west, with one exception. A plane which strikes absolutely east—west is given in the conventional manner an angle of strike of $+90^\circ$. Such a plane can dip to the north or to the south. The first case, with dip to the north, is referred to as dip to the east, while the opposite case, with dip to the south is referred to as dip to the west. For purposes of simplicity dip to the west is referred to as positive and dip to the east as negative, in a similar manner to that used with strike. Thus dip also becomes an angle varying between -90° and $+90^\circ$.

16.2 The description of the direction of lines in space

The angle of strike of a line is defined as the angle between a vertical plane passing through the line and the direction north—south. Dip is the

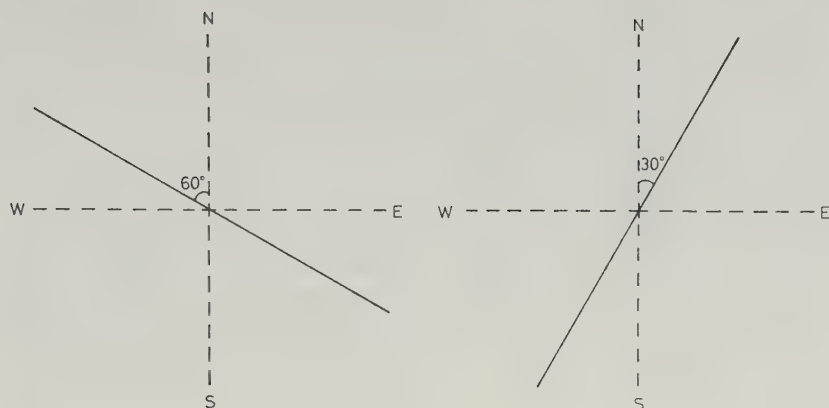


Figure 16.3. The strike of a plane is defined as the angle between a north-south line and the line of intersection between the plane and the horizontal plane. The angle of strike is $N60^{\circ}W$ in the left hand diagram and $N30^{\circ}E$ in the right hand diagram.

angle between the line and the horizontal plane. These angles are referred to in precisely the same way as in the case of the plane.

16.3 Fabric diagram — planes

A fabric diagram is constructed in the following way. One imagines one has a half sphere, actually that half sphere that lies below the horizontal plane. Also one has a plane, the direction of which is to be described in the diagram. One positions the half sphere such that the plane goes through the centre point of the sphere. See figure 16.4. In this figure the whole plane is not shown, but only a small part of it, a circular area which represents the whole plane. Due to the fact that the circle is seen diagonally from the side it is drawn in the form of a shaded ellipse. A normal to the plane is then drawn from the centre of the sphere to a point on its outer surface, which is referred to as point Sp.

In figure 16.5 the sphere and the circle are seen from directly overhead, i.e. from point B in figure 16.4. Even in this case the circle is seen as an ellipse, the main axis of which is a part of the line of intersection between the horizontal plane and the plane in question. One can also see the normal to the plane and the point of intersection between the normal and the surface of the sphere which is referred to as Sp. From this figure it can be seen that the angle of strike (*strike*) is approximately $N65^{\circ}E$.

The construction of a line between point A in figure 16.4 and the centre of the sphere gives the line of intersection between the plane and the

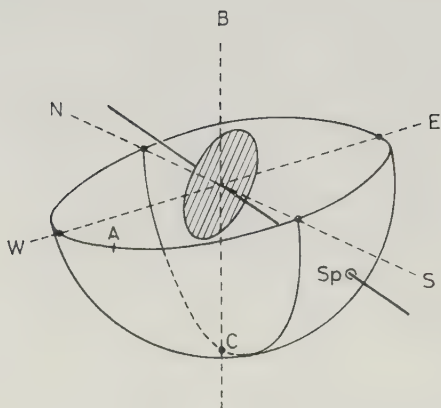


Figure 16.4. A small circular area has been placed with its centre point at the centre of the half sphere. A normal to the circular area is then drawn from this centre point and this normal intersects with surface of the half sphere at point Sp.

horizontal plane. Figure 16.6 shows the half sphere and the circle from a point on this line. In this case one only sees the diameter of the circle. Included in the figure are the normal to the plane and the point Sp. The angle of dip is about 60° , i.e. the plane dips about 60° to the northwest. The direction of the plane is thus $(-65^\circ, 60^\circ)$ using the convention given in section 16.1. The intersection between the normal and the surface of the sphere describes completely the angles of strike and dip. This is a highly economical way of describing the direction taken by a plane.

The procedure described in figures 16.4 to 16.6 is then repeated for the next plane and so on, until all the planes have been treated. The result is a half sphere with a number of points on its surface, one point for every plane. It is sometimes possible to see where on the surface of the sphere, the greatest concentration of points is to be found, giving indirectly the dominating direction amongst the planes. However, in most cases an isarithmic diagram (a fabric diagram) is constructed showing the number of points per reference area, where the reference area is taken to be a circle having an area 1% of the surface area of the half sphere.

16.4 Fabric diagram — lines

Even in cases where the directions taken by a number of lines in space are to be analysed the fabric diagram method can be used. The lines are assumed to pass through the centre of the half sphere and intersect with the surface of the sphere at a point. Thus, in the case of lines, the line itself is used and not the normal. When all the lines have been treated the construction of the diagram can proceed in the same manner as for planes. The difference between the construction of a fabric diagram using lines, as

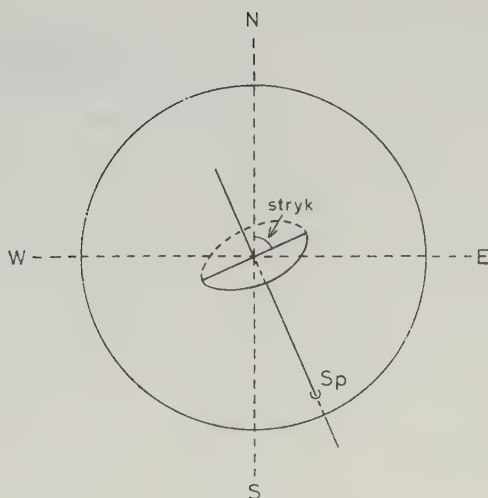


Figure 16.5. The half sphere seen from point B in figure 16.4. The angle of strike is denoted by *stryk*.

opposed to planes, is that in the first case the normal to the plane is used and in the second case the line itself. Thus, care must be taken when giving angles of strike and dip in order to distinguish between lines and planes.

16.5 The projection of the half sphere on to an xy plane

The 3-dimensional fabric diagram described in section 16.3 is too complicated for practical use. The method usually used is to project the surface of the half sphere on to the xy plane. This is usually done using Lambert's equal area map projection. The result is a so-called Schmidt's net as shown in figure 16.7. Such a net has the same total area as has the surface of the corresponding half sphere. Use of a stereographic projection gives a directionally correct net called Wulff's net.

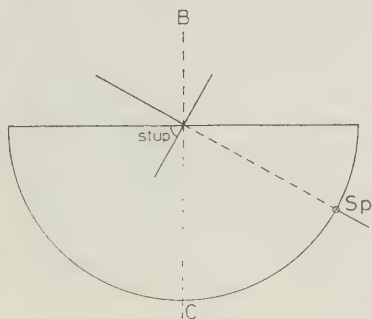


Figure 16.6. The half sphere seen from point A in figure 16.4. The angle of dip is denoted by *stup*.

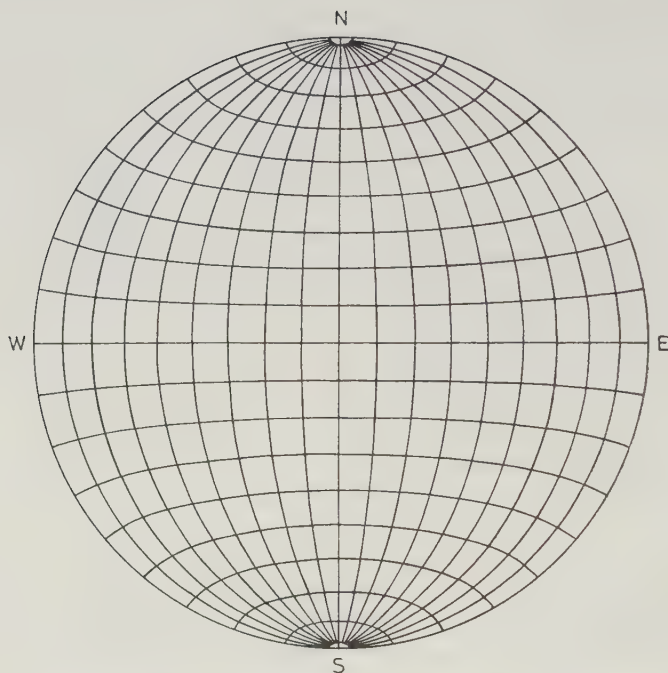


Figure 16.7. Schmidt's net, i.e. Lambert's equal area equatorial projection of a sphere.

The net shown in figure 16.7 results from the choice of a point on the equator of the sphere as the centre of projection. This can be seen from the fact that the meridians appear as curved lines about the WE axis. This net is only indirectly used when constructing a fabric diagram of the type described here (line or axis fabric diagram). However, it is used when one has very few sedimentary or tectonic planes and wants to describe them by aid of a fabric diagram. In such cases the whole intersection between the surface and the half sphere is shown in the diagram, producing another type of fabric diagram, which can be called a surface fabric diagram.

In reality the net used in the construction of a fabric diagram has the point C in figure 16.4 as the centre of projection. The surface of the sphere is graduated in degrees, where the lines of latitude are graduated with point C at 0° and the equator at 90° . The lines of longitude are graduated from point E in the same way. Point N become $+90^\circ$ and point S -90° . The surface of the sphere, together with its net of graduations, is projected on to a plane using Lambert's equal area map projection, giving a result of the type shown in figure 16.8. The use of this net simplifies both the construction and the use of the fabric diagram, due to the fact that the angles of strike and dip can be read directly.

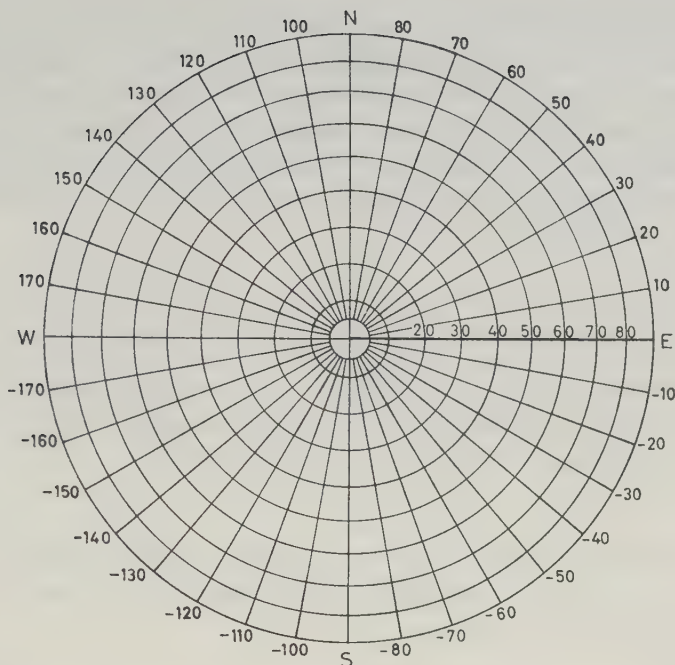


Figure 16.8. The net obtained when using the pole of the sphere as the centre of projection in conjunction with Lambert's equal area projection.

The calculations necessary in order to carry out the projection are done in the following way. The polar coordinates of the point of intersection Sp (r, fi), or more correctly of the projection of the point in Schmidt's net are calculated on the basis of the angles *stryk* and *stup* which are the angles of strike and dip. As can be seen from figure 16.9 r is the distance between point Sp and the centre of the net at point C , and fi is the angle between the base line CE and the line CSp .

The r coordinate is dependent only on the angle of dip and is calculated using formula 16.1. In the formula *stup* is taken to refer to the angle of dip. The angle fi is dependent on both angle of strike and angle of dip,

$$r = abs(100 \sqrt{2} \sin(stup/2)) \quad (16.1)$$

dependent on whether the plane dips to the west (dip larger than 0) or to the east (dip equal to or less than 0). The angle fi is calculated using formula 16.2, where *stryk* refers to the angle of strike. Take special notice of the relationship $-90^\circ < stryk \leq +90^\circ$.

$$\begin{aligned} 1. \quad fi &= stryk & \text{if } stup > 0 \\ 2. \quad fi &= stryk + 180^\circ & \text{if } stup \leq 0 \end{aligned} \quad (16.2)$$

The calculations using formulas 16.1 and 16.2 give as their result the polar coordinates (r, fi) in the polar coordinate system shown in figure 16.9. If it is necessary to use a dimension other than 100 mm for the diameter of the net the new dimension is used instead of 100 in formula 16.1. The polar coordinates (r, fi) are then transformed to xy coordinates, in order to provide the input data to the isarithm program. This transformation is done using formula 16.3.

$$\begin{aligned} x &= r \cos(fi) \\ y &= r \sin(fi) \end{aligned} \quad (16.3)$$

The result of this operation is the xy coordinates for a number of points. The NORI program can then be used to give an isarithmic diagram. The function mapped with the help of the fabric diagram is the number of points Sp per reference circle having a radius of 10 mm.

16.6 The projection on to the xy plane of the point of intersection between a line (b -axis) and the surface of the half sphere

In cases when a line, a b -axis for example, is used the polar coordinates of its point of intersection with the surface of the sphere (r, fi) , are required. Formula 16.1 can be used to calculate r , but a new formula is required for fi . One can distinguish between two cases. In the first the dip is not in the same direction as the strike. If the line strikes to the east then it must dip to the west and vice versa. As dip ($stup$) and strike ($stryk$) are given using the signs + and - this restriction can be stated most simply as $stup \times stryk$ should be less than 0. In such a case one adds 270° to the angle of strike in order to obtain fi . In the second case where strike and

$$\begin{aligned} 1. \quad fi &= stryk + 270^\circ \text{ if } stup \times stryk < 0 \\ 2. \quad fi &= stryk + 90^\circ \text{ if } stup \times stryk \geq 0 \end{aligned} \quad (16.4)$$

dip have the same sign the angle fi is obtained by adding 90° to the angle of strike. The two cases are given in formula 16.4. The polar coordinates are then transformed to xy coordinates using formula 16.3.

The result obtained when using data for a line is also a point in the plane xy , which can then be processed using NORI.

16.7 Values at the periphery of the net

The line of intersection between the horizontal plane and the half sphere of the fabric diagram can, for the sake of simplicity, be called the equator

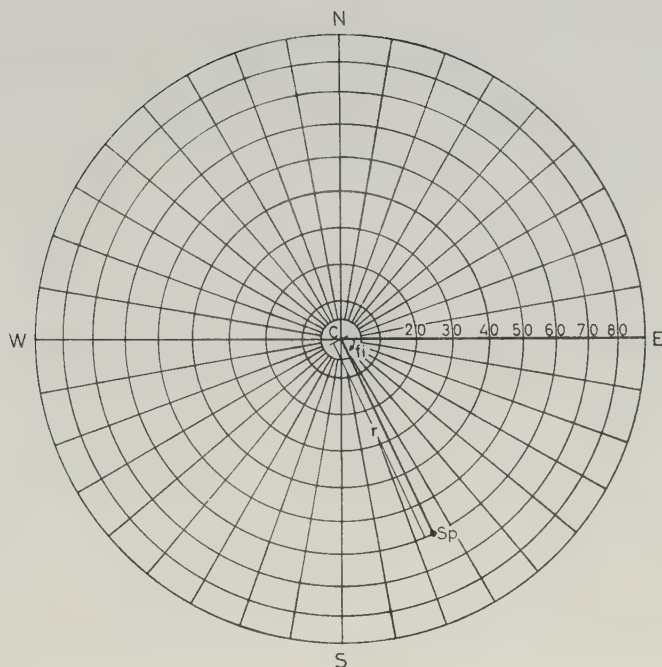


Figure 16.9. The distance between the centre of the net, C , and the projection of the point Sp on to the net is referred to as r . Sp is the point of intersection between the normal to the plane and the surface of the sphere. The angle between the line CE and the normal is referred to as fi .

of the sphere. A reference area lying on or just under this equator will lie partly over the horizontal plane. As the upper half of the sphere is not included when formulas 16.1 to 16.4 are applied a deformed reference circle is obtained. This has the effect that the value of the function at the points on the equator is only half what it should be.

All reference areas the centres of which lie less than 10 mm from the equator are deformed to some degree, which results in values less than the correct value of the function. Every plane should make equal contributions when the value of the function is being calculated irrespective of how steeply the plane dips. It is therefore desirable to correct, as far as possible, this systematic error. When calculating the number of intersection points Sp per reference area regard should be paid to that part of the reference area lying over the horizontal plane. The simplest way of doing this is to calculate the intersection point between the normal and the surface of the sphere above the horizontal plane. Thus two intersection points are cal-

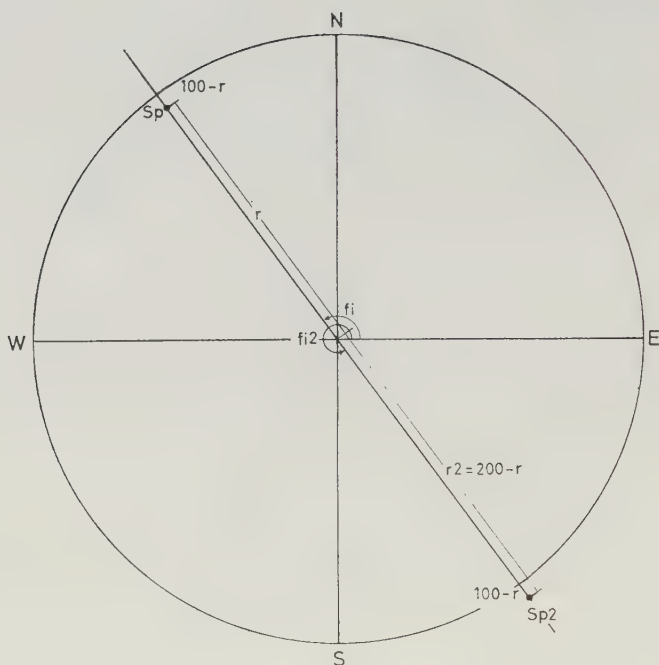


Figure 16.10. If a point Sp having the polar coordinates (r, fi) lies in the vicinity of the periphery of the net the position of a further point $Sp2$ having the polar coordinates $(r2, fi2)$ must be determined. $Sp2$ is the point of intersection between the normal to the plane and the upper half sphere.

culated for every plane; Sp which lies on the lower half of the sphere and $Sp2$ which lies on the upper half of the sphere. The distance between the equator and point Sp needs to be less than the radius of the reference circle before it is necessary to determine the position of point $Sp2$.

The periphery of the net is equivalent to the equator of the sphere. Also points Sp and $Sp2$ lie symmetrically on either side of this equator. If the distance from point C to point Sp is given as r the distance from Sp to the periphery is $(100-r)$ mm. See figure 16.10. The distance between point Sp and $Sp2$ is always equal to the diameter of the net, in this case 200 mm. As a result the distance $r2$ between C and $Sp2$ is equal to $(200-r)$ mm. See figure 16.10. Also 180° must be added to the angle fi to give the angle $fi2$. The point $Sp2$ has thus the polar coordinates $(r2, fi2)$. Formula 16.4 can then be used to transform the polar coordinates (r, fi) and $(r2, fi2)$ to xy coordinates (x, y) and $(x2, y2)$. Formula 16.5 summarizes the whole procedure.

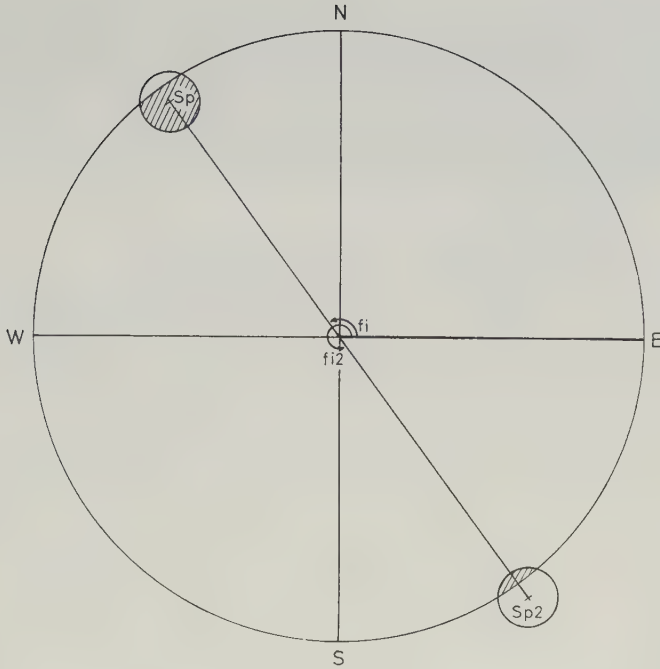


Figure 16.11. An influence area placed over point Sp when this point lies near the periphery of the net is deformed and requires correction using an influence area having point $Sp2$ as its centre.

$$\begin{cases} r2 = 200 - r \\ fi2 = fi + 180^\circ \\ \begin{cases} x = r \cos(fi) \\ y = r \sin(fi) \end{cases} \\ \begin{cases} x2 = r2 \cos(fi2) = (200 - r) \cos(fi + 180^\circ) \\ y2 = r2 \sin(fi2) = (200 - r) \sin(fi + 180^\circ) \end{cases} \end{cases} \quad (16.5)$$

The NORI program works in the following manner; an influence area is created around every point Sp or $Sp2$. The value of the function at all the grid points lying within the influence area is then increased by the quantity of data associated with point (x,y) in this case the value 1. When a point Sp lies near the periphery of the net the influence area is deformed, as is the influence area around point $Sp2$. As can be seen from figure 16.11 the sum of the two deformed influence areas is the same as one non-deformed influence area. As long as one limits oneself to grid points lying inside the net correct results are obtained. This is also the case with points lying just outside the periphery of the net. The further the point lies outside the periphery of the net the more inaccurate the result becomes. This

increasing inaccuracy is due to the fact that the projection is only correct for areas within the periphery of the net. The inaccuracy which occurs when using points lying just outside the periphery of the net is more than compensated for the contribution they make to increasing the accuracy of the values associated with points lying immediately within the periphery.

16.8 The fabric diagram program NORIGE

The addition of a fabric diagram section to the NORI program creates the NORIGE program. The procedure call contains the following parameters $X0$, $Y0$, L , W , H , MAP , $INUNIT$, $ITYPE$ and $IDATA$. The parameters $X0$, $Y0$, L , W and H have the same meaning as in NORI, i.e. $(X0, Y0)$ is the origin of the diagram, L its length, W its width and H the equidistance. These parameters can be given fixed values in the program, as can the overlapping constant, thus avoiding having to read them in every time the program is used. In such a case one can use the standard dimension for Schmidt's net of 100 mm. The reference area is then always a circle having a radius of 10 mm. The overlapping constant can be given as 5.0, which results in a mean error in the position of the isarithms of $0.07 H$ (Nordbeck & Rystedt 1970). As H in this case is 4 mm the mean inaccuracy is about 0.3 mm. If this is considered too large the value of S can be given as 10.0 and H as 2, resulting in a mean error of 0.1 mm. It is best if the central point in Schmidt's net is a grid point. When H is equal to 4 mm L and W are equal to 51, 25 rows above the central point and 25 below, also, in the same manner, 25 columns to the left of the central point and 25 columns to the right. When H is equal to 2 mm L and W are equal to 101. The overlapping constant S should then be allocated the value 10.0 giving a reference area with a diameter of 20 mm. The origin of the fabric diagram has the coordinates $(X0, Y0)$. As a result of using the radius 100 mm for Schmidt's net the values of $X0$ and $Y0$ are equal to -100. The units of the coordinate system are in this case mm.

The parameters $INUNIT$, $ITYPE$ and $IDATA$ describe where the data are to be obtained and how they are to be handled. $ITYPE$ equal to 1 indicates that lines are being processed, $INUNIT$ refers to the logical unit number used for input. If $INUNIT$ is equal to 5 the data describing strike (*stryke*) and dip (*stup*) is read from punched cards. If $IDATA$ is 0 a further variable $WGHT$ is read from $INUNIT$ together with every plane's angles of strike and dip. A value other than 0 for $IDATA$ results in the value chosen being assigned to every plane. The variable $WGHT$ can be treated as a weighting, assigned to the plane in question according to given rules.

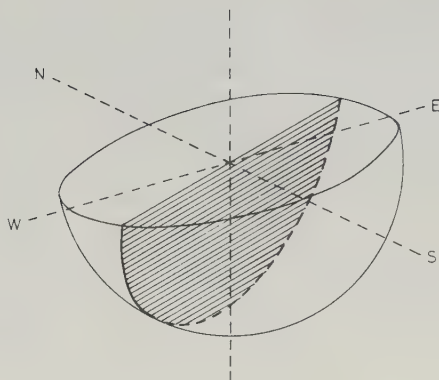


Figure 16.12. When an area fabric diagram is being prepared the whole intersection line between the plane and the surface of the sphere is entered into the diagram.

A higher value can be given to *WGHT* if the plane in question is large and therefore more important than smaller planes.

16.9 Area fabric diagrams and NORYG

In those cases where sedimentary or shale surfaces are to be shown (so called *s*-areas) a different technique is used. Instead of the axis fabric diagram one obtains the area fabric diagram. In this type of fabric diagram the whole line of intersection between the plane and the half sphere is shown, see figure 16.12.

One obtains a circle for each plane studied. The points of intersection between these circles on the surface of the half sphere can be recorded and their distribution studied. In this way it can be decided whether one is dealing with one or more systems of approximately parallel planes, together with the direction taken by forces responsible for deformations. Of course, it is impractical to work with the half sphere in 3 dimensions, therefore it is projected on to a plane, together with the points of intersection of the circles, in the way described in section 16.5.

As can be seen from figure 16.12 the intersection line between a plane and the half sphere is always half a great circle. It is still an arc when projected from the sphere to Schmidt's net. The circle, of which the projected intersection line forms a part, is shown in figure 16.13 as having a radius r_s , with its centre lying at a point having the polar coordinates (r_3, f_i) . In this case f_i is defined in accordance with formula 16.2. One calculates r_s by aid of formula 16.6 and r_3 by aid of formula 16.7. The polar coordinates (r_3, f_i) , can be converted to xy coordinates by the use of formula 16.3 giving the coordinates (x_3, y_3) .

$$r_s = 25\sqrt{2} (2 - \sin(\text{abs}(\text{stup}))/\sin(45^\circ - \text{abs}(\text{stup}/2))) \quad (16.6)$$

$$r_3 = 25\sqrt{2} \sin(\text{abs}(\text{stup}))/\sin(45^\circ - \text{abs}(\text{stup}/2)) \quad (16.7)$$

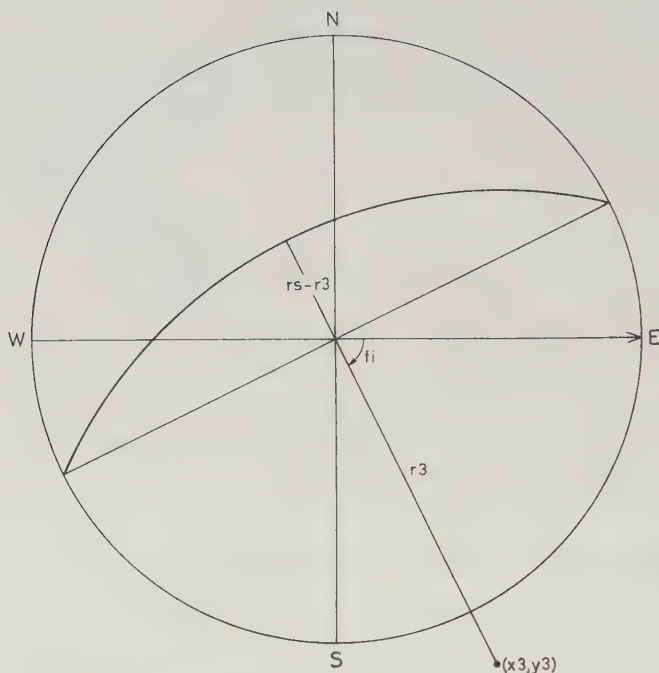


Figure 16.13. The intersection line between the plane and the surface of the sphere projected on to Schmidt's net. The line takes the form of a circle segment, the middle point of the corresponding circle lies at a point having the polar coordinates (r_3, f_i) and its radius is equal to r_s .

In that a circle's radius and the location of its centre are known its equation is also known. When two such equations are available it is a simple operation to determine the location of any possible points of intersection between the two circles. These points of intersection can then be used as input data to another program, for example NORI, and an isarithmic diagram showing their distribution constructed.

The NORYG procedure is used when constructing a fabric diagram of the type described above. The procedure has exactly the same parameters, variables and indata as the NORIGE procedure, with the exception of the parameter *IDATA* and the variable *WGHT*. The procedure results in a numerical map, suitable for further processing to give an isarithmic map.

16.10 The drawing of the fabric diagram and NORGE

As was mentioned in section 16.9 it is often desirable to draw out the intersection between the half sphere and the *s*-areas or, more correctly, the

projection of the half circles in Schmidt's net. Also it is often desirable to indicate the point of intersection between the half sphere and the normal to the plane (the pole of the plane). The NORGEP program determines the location of the intersection lines and the poles and draws them as an area fabric diagram. The poles are marked in this diagram by a cross, while the intersection lines themselves are drawn as arcs.

The NORGEP procedure contains the same parameters and variables as NORYG. In addition it contains the parameter *DRAW* and the variable *THETA*. *THETA* is the size of the arc to be drawn given as an angle. *DRAW* is used to indicate whether one wishes to show only the arcs, ($DRAW = 2$), only the poles, ($DRAW = 1$), or both, ($DRAW = 0$).

17 The main program of the NORMAP system

As has already been pointed out in chapter 1 NORMAP consists of a number of procedures, or subroutines, which are fairly simple to combine into a main program. How this is done is best shown using an example. A prerequisite is that data is available in the form of a square net map and that this data is to be processed to give either a new square net map or an isarithmic map. The whole of this process can be carried out completely automatically, including both the construction and the drawing out of the map.

17.1 Data

Data for the example is provided by the square net map of population distribution in southern Ghana prepared by Claeson, Godlund and Örtengren (1969). See figure 12.1. The map is read into the computer by procedure MAPIN, which then transfers the data to a data file. The data should be available on punched cards in the form, $X, Y, Z_X, Z_{X+1}, Z_{X+2}, \dots$. In this case Z_{X+N} refers to the population of the grid square having coordinates $(X+N, Y)$. This means that the map is read in row by row. Every punched card must begin with an X coordinate, which means that long rows occupying several punched cards require an X coordinate on every card. The special case $N=0$ occurs fairly often, but even in this case MAPIN can be used although each card only contains information about one point. If one is using a register containing a number of different items of data, referring to the same point, a special procedure NORXRP is required to extract the relevant Z value from the register and to carry out the processing necessary to produce a magnetic tape of the MAPIN type, with the data arranged as triples (X, Y, Z) .


```
VRUN ISPLOT,ACCONT,PROJECT,1,30
VHDG AN EXAMPLE ON A MAIN PROGRAM USING SUBROUTINES IN NORMAP
VASG,A NOR*NORMAP.,F2
VASG,T 8.,F2
VASG,UP PLOT*GHANA.,F2
VFOR,IS ISPLOT
    DIMENSION MAP(18,36)
    CALL MAPIN(8)
    CALL NORK(0,0,18,36,1,MAP,8)
    REWIND 8
    CALL MAPOUT(0,0,18,36,1,MAP,4)
    CALL NORI(0,0,10,19,2,MAP,4,3,0,8)
    CALL MAPOUT(0,0,10,19,2,MAP,3)
    CALL NORIP(0,0,10,19,2,MAP,3,0,6,1341119.,1.0)
    END
VMAP,IS ISPLOT
IN ISPLOT
LIB NOR*NORMAP.
VBRKPT PRINT$/PLOT*GHANA
VXQT ISPLOT
24
(214,2413)
2 17 1 1 1 8 8 11 8 9 6 7 19 12 24 22 2 1
23 17 1 2 2 6 21 9 5
1 16 1 1 1 1 9 3 3 5 9 9 21 36 24 19 4
17 16 13 2 1 1 4 16 18 9
0 15 1 1 2 1 7 7 5 1 10 16123123 29 4
16 15 12 10 4 14 14 4 5 1 9 11 6 1
1 14 1 4 1 1 2 6 4 3 11 9 10 16 20 13 12 19
17 14 17 15 36 18 8 7 1 6 3 2 19 11 21 3 6
2 13 3 4 8 6 3 22 2 3 12 11 9 26 9 13 7 4
18 13 8 6 3 11 2 3 27 8 4 9 5 1 2 7 2
2 12 4 4 5 3 7 4 3 5 17 12 7 6 1 9
18 12 7 7 9 14 16 18 18 8 6 7 1 1 7 5 9 6 1
2 11 6 4 2 2 6 13 5 5 2 4 32 8 6 3 9 9
18 11 6 11 9 13 22 22 21 15 20 4 2 3 3 5 9 20
34 11 17 13
1 10 2 1 1 1 4 3 2 4 6 2 5 1 10
17 10 4 5 35 8 14 23 70 14 16 9 9 11 12 11 6 17
33 10 19 24 10
7 9 3 1 2 4 19 3 2 4 6 4 21 20 14 8 15 14
23 9 11 16 30 6 4 17 11 17 40
3 8 7 1 2 10 1 1 5 1 3 1 4 2 3 20 21
19 8 11 29 25 37 21 14 3 1 3 13 9 18 15 10 23
3 7 2 4 9 5 1 4 21 4 8 7 5 18
19 7 28 10 23 6 10 30 20 20
4 6 1 5 1 1 3 6 3 1 7 5 1 5 14
18 6 4 25 39 15 3 7260100
4 5 2 1 21 3 3
13 5 5 11 1 13 12 49 31 9 12 13 2
4 4 4 42 10 1
13 4 1 8 4 14 12 16 26 13 11
8 3 1 8 3 1 2 6 9 6 20 16 16 23
1 2 2 8 8 16 7 8 2 3 3 5 4 20 15 17 70 20
5 1 8 6 10 5 4 11100 50
8 0 9 4 9 3
-1
9 100 200 300 400 500 600 700 800 900
V FIN
VRUN,/S ISPLOT,ACCONT,PROJECT,1,30
VSYM PLOT*GHANA.,PR6
V FIN
```

17.2 The main program ISPLOT

The ISPLOT program is written to be used on the Univac 1108 computer at the Lund Computer Centre. Thus, the conventions used at this centre occur in the program shown here. It is a fairly straightforward matter to replace these conventions with those used by other computers and other computing centres.

Card 1 contains the control statement RUN. It has a 7-punching in the first column and the word RUN in columns 2 to 4. This is followed by a space and the name of the program, ISPLOT, followed by a comma. ACCOUNT and PROJECT stand for the number of the account and the name of the project which is to bear the cost of the operation. The number 1 refers to the fact that the program will require less than 1 minute for its execution. In the same manner 30 refers to the maximum number of pages which are to be printed on the line printer. The 5 groups, program name, account number, project, maximum time and maximum pages are separated by commas.

Card 2 is a heading card. The text after HDG specifies a heading to be printed at the top of each page of printed output, along with the page number and the current date. The next three cards assign mass storage facilities to this run, under the specified file names. The file with the name NOR*NORMAP contains the NORMAP program package. On file 8 the result from MAPIN is to be stored, and on file PLOT*GHANA the plotted map is to be stored for later output to the line plotter.

Card 8 contains a call on the MAPIN procedure, which reads in the square net map and stores it on the mass storage file logical unit number 8.

Card 9 contains a call on the NORK procedure. This procedure constructs a numerical map having its origin at point (0,0). The map has 18 rows and 36 columns and the equidistance is 1. The input data is obtained from the logical unit number 8.

Card 11 calls MAPOUT which prints out the map so obtained. In this case the parameter *GFORM* has the value 4 which means that MAPOUT is to use a quadratic grid net giving the result in the form of a square net map.

Card 12 calls the NORI procedure. This procedure constructs a map having its origin at point (0,0) containing 10 rows and 19 columns of grid points. In this case the equidistance is 2. The NORI map covers the same area as the square net map shown in figure 12.1. The NORI map is stored in the array *MAP*. The reference areas are square, which means that

RFORM must be given the value 4. The overlapping constant is 3.0 and the input data is available via logical unit 8.

Card 13 calls the *MAPOUT* procedure, which is used to output the map constructed using *NORI*. This means that the first 6 parameters in the procedure call should be identical with the parameters used in the call on the *NORI* procedure given on card 12. These 6 parameters describe the location of the map, its size, equidistance and where it is stored. The procedure call also contains a parameter *GFORM* which describes the form to be taken by the grid net. The grid net in the example is triangular, as the net is triangular the value 3 is used. Card 14 calls the *NORIP* procedure. The first 6 parameters in the call are the parameters of the map, i.e. the same as those used in the call on *NORI* (card 12) and the call on *MAPOUT* (card 13). The *NORIP* procedure interpolates the position of the isarithms using a triangular grid net. *GFORM* must therefore be 3. The parameter *LUN* is given the value 0, which means that the isarithm polygons are not stored but directly drawn on the plotter. The next parameter refers to the grid which is to be drawn by the plotter at the same time as the isarithms. The value 6 for this parameter results in a grid having sides 50 miles long on the ground. The units used for the coordinates, in this case 8.333... miles should be given in centimetres. As 8.333... miles is equal to 1 341 119 centimetres *CUNIT* is given this value. 1.0 refers to the length of the unit used for the coordinates in centimetres on the finished map.

With card 16 the processor *MAP* is called to form an executable program. In this program the main program *ISPLOT* and the subroutines in the library file *NOR* NORMAP* are included. The control statement *BRKPT* switches over the output from the line printer to the file *PLOT*GHANA*.

Card 21 gives the largest number of *Z* values occurring on a data card. In the example this is 24.

Card 22 gives the format used for the data cards. First 2 integers of 4 figures each. These are the *X* and *Y* coordinates of the first grid square in the row. They are followed by 24 integers of 3 figures each. These 24 are the *Z* values, the first of them being the population of the grid square described by the *X* and *Y* coordinates; those following being the values for succeeding grid squares in the same row. The grid squares are dealt with from left to right.

Card 23 and the succeeding cards are data cards punched using the format given on card 22. The last data card should have a negative *X* coordinate, in this case -1. The computer is informed that all data cards have been read in when it comes across a card having a negative *X* coordinate; the

input of data cards is then discontinued and the processing of the data begun.

The next card gives the number of isarithms required and the levels at which they are required. In the example 10 integers of 5 figures each are shown. The first of these is the number of isarithms, the positions of which are to be determined and drawn out. The following 9 number are the values of the different isarithms, in this case one obtains an isarithmic map showing the 100, 200, 300, 400, 500, 600, 700, 800 and 900 isarithms. The unit is in this case population in thousands per 2500 miles².

The last three cards are a run which outputs the file PLOT*GHANA to the output device PR6. This device contains both a line printer and a line plotter. The resulting map is shown in figure 6.4, page 57.

The use of only about 20 program cards has produced a program capable of carrying out a relatively complicated processing of the given data. ISPLOT reads in a square net map, constructs a new square net map and prints it out. It constructs a map of population density using the overlapping reference area technique and a triangular grid net and prints it out. Then the program interpolates the location of 9 isarithms, 100, . . . , 900 thousands of persons per 2500/miles². Finally the isarithms are drawn in to give an isarithmic map, with a graduated grid superimposed and with the values of the isarithms printed against each isarithm. The actual calculations are carried out exceedingly quickly, 6 seconds were required for the Ghana example given above when using the Univac 1108 computer in Lund.

Part II

Location models

1 The spatial pattern of central places

If one studies a map of a sufficiently large area, such as a whole country, one finds that its towns and built-up areas are located according to a more or less regular pattern. Kant (1951) describes this pattern as being circular with the largest town at the centre of a pattern of concentric rings. In the ring lying nearest the central place only small and unimportant built-up areas are found. Further out another ring of somewhat larger towns or built-up areas is to be found. Around the central places in the outer ring the same pattern repeats itself.

Kant developed his model on the basis of the pattern of built-up areas in Estonia in the 1930's. The correspondence between the model and the actual pattern was, in this case, not absolute. These deviations can be partly explained by the distribution of population, the communication net and the terrain.

In Skåne, a part of southern Sweden, Malmö is the central place. Application of Kant's model to this region would indicate a ring of very small central places around Malmö, as is actually the case. A number of these places function as dormitory suburbs to Malmö and have, therefore, a larger population than that predicted by the model. A built-up area of this type should probably be considered to be a part of Malmö and not as a central place in its own right. Looked at from a functional point of view Malmö and Lund form a single large central place. When this assumption is made Kant's model applies to Skåne exceptionally well.

In Skåne there is first a ring containing the towns Landskrona, Eslöv, Höör, Hörby, Sjöbo, Skurup and Trelleborg. Outside this ring is another containing larger towns than those of the first ring. Hence, Helsingborg is larger than Landskrona, Kristianstad is larger than Hörby and Ystad is larger than Skurup.

Kant (1951) divides central places into 5 categories, large towns, towns, small towns, dwarf towns and town-like built-up areas in country districts. He thus uses a hierarchy with 5 levels. Christaller (1933) who was one of

the pioneers in this subject used 7 levels: Landeshauptstadt (L), Provinzhauptstadt (P), Gaustadt (G), Bezirkstadt (B), Kreisstadt (K), Amtsort (A) and Marktort (M). A Landeshauptstadt is the most highly developed central place, with the highest degree of centrality, and Marktort lies at the lowest level in the hierarchy. Godlund (1951) uses the terms regional centre, district centre, neighbourhood centre and local centre.

Central places at the same level in a hierarchy are located in a regular pattern. Christaller (1933) found that every central place was located at a node in a triangular net and that the corresponding umland was a regular hexagon. Other authors have found that a model using regular pentagons or irregular polygons is more in accordance with conditions in their particular area (Hannerberg 1961). Godlund (1951) constructed a map showing the size of the umlands corresponding to the regional centres in Skåne, also a similar map showing the bus umlands. The majority of the umlands he shows are irregular pentagons or hexagons.

A large town is more important to its surroundings than a small town. This effect is referred to here as centrality. Godlund (1956) calculates centrality as a function of the shopping turnover, amongst other factors. The centrality index, thus arrived at, is used by him when defining the umland a particular central place is considered to have as against other central places in the district. The central place theory and Godlund's application of it, i.e. his technique for defining umlands has been widely used by Swedish planners. It was used as an aid in the reform of local authority boundaries (Jakobsson 1964). This reform aimed to combine local authorities to produce financially viable groupings. The number of local authorities in Sweden has been reduced, as a result of this reform, from about 1000 in 1965 to about 280 in 1974.

1.1 Central place theories

Christaller (1933) bases his central place theory on a basic unit of settlement. He makes 4 basic assumptions that:

1. Population is evenly distributed over the surface.
2. The transport surface is homogeneous, i.e. it is equally easy to transport goods in all directions.
3. The basic units of settlement are located at the nodes of a triangular net consisting of triangles having sides of equal length.
4. All umlands are hexagonal.

When specialization occurs e.g. trading, meaning that more than one basic unit is necessary to provide sufficient customers for a particular facility,

then some form of grouping of basic units is necessary. This grouping occurs around a central unit. Christaller assumes that a fixed number of basic units are necessary in every group in order to support a particular facility. This fixed number is referred to as k .

In the special case when k is 7 the grouping is simply understood. The umland corresponding to the central unit is expanded to include the 6 basic units lying nearest the central unit. When k is 3 every basic unit is distributed between three different central units, k equal to 4 means that a basic unit is distributed between two different central units.

It will be noticed that k cannot assume some values e.g. 5 due to the fact that it is, in these cases, impossible to fit a hexagonal pattern to the given distribution of basic units.

The same principles can then be applied to the central units as were applied to the basic units, giving central places at the next higher level in the hierarchy. At the highest level there is only 1 central place. The next highest level has k central places, at the third level there are k^2 central places and so on. A geometric series is obtained i.e. $1, k, k^2, k^3, \dots$

Lösch (1954) uses a considerably more complicated central place model than Christaller, but starts from the same basic assumptions. He also assumes that transport costs increase as distance increases thus making the price of an item higher the longer one travels from the central place. As a result, the demand for an item is lower at the periphery of a particular umland than nearer to the central place. He then constructs a 3-dimensional demand diagram and the result is what he refers to as a "demand cone". He then attempts to fit these demand cones to the area in question, with the restriction that the whole area is covered using the cone bases without overlapping. He discovered that the cone bases needed to be regular hexagons. The demand is maximized when the umland is a regular hexagon. Lösch has a businessman's point of view and states that all location models must include profit maximization.

According to Lösch every service facility has its own pattern of location, with the central units arranged at the nodes of a triangular net and the umlands forming regular hexagons. One then chooses a facility lying near the centre of the pattern and call this point the centre of the pattern. Lösch then places the different patterns and their central facilities over one another in such a way that the largest possible number of facilities coincide. This results in 6 sectors with a large number of central places and 6 with only a small number of central places. The principle is that the largest central place contains an example of every service facility, while outside of the largest central place the facilities are located in such a way

that the range of facilities is as rich as possible. Lösch includes Christaller's fixed k model as a special case of his own model.

Amongst others that have worked with central place models should be mentioned Ullman (1941) who applied the model to American conditions. Berry and Garrison (1958) gave a clearer definition of the terms "threshold level" and "the range of a central good". They defined "the threshold level of a good as the minimum sales level required in a central place to make it worth providing the good and the range of a good as the radial distance which delineates the market area of a central place for the good". The umlands they obtain are hexagonal. Vidale (1956) has constructed a model which gives hexagonal umlands when Christaller's basic assumptions are made. If, on the other hand, the assumption is made that population is not evenly distributed over the surface the hexagonal umlands are not obtained.

1.2 The hierarchy of central places and the rank size rule

Christaller's central place model is characterized by the fact that the central places are arranged according to a hierarchical system and by the fact that all central places at the same level in the hierarchy are equally large. Christaller himself discovered many deviations, when applying his model to the towns of southern Germany. He stated that the most likely explanation for these deviations was the fact that the transport surface was not homogeneous. Vining (1955) doubts that central places are organized according to a formal hierarchical system. He also doubts that umlands have a particular geometric form.

In order to discover where a central place falls in the hierarchy its centrality must be determined. Christaller did this by using the quantity of telephone traffic occurring within the town, as against the quantity of traffic directed to other places outside the town. A simpler method is the construction of lists describing the activities which are assumed to be present at a particular level in the hierarchy. The beginning of such a list could look like this: Government, University, County authority. Up to a number of years ago the Swedish lists resulted in the numbers of towns at each level in the hierarchy forming an almost perfect series, 1, 4, 19. In the last few years new universities have been founded which means that new activities must be added to the lists if the perfect series is to be retained. These lists contain available activities and goods arranged in such a way that a hierarchical arrangement of central places is produced.

This list method gives no proof that a formal central place hierarchy exists. Instead it is based on assumption that the hierarchy exists.

If central places are arranged in order of precedence on the basis of population, giving the largest the number 1, the next largest the number 2 and so on one finds that population multiplied by the order of precedence is a constant C . If the population of a central place is P and its rank number is n the expression $n \cdot P = C$ is true. C is thus approximately equal to the population of the largest central place.

This rank size rule is described by Zipf (1949) amongst others and can be easily proved using the law of allometric growth (Nordbeck 1965, 1971). It has been verified using actual data in a large number of different investigations. It can be directly seen that the rank size rule contradicts the central place theory with its similarly sized central places arranged at the same level in the hierarchy. Beckman (1958) attempts to prove that both models complement one another. This is done by introducing a random element into the central place theory. Bunge (1966) claims that the rank size rule is the most important proof of Christaller's urban hierarchy. If one cannot accept Beckman's ideas regarding the presence of a random element in the central place theory then the rank size rule must prevent an acceptance of the central place theory. This contradiction between the two models can be established without the necessity for any decision regarding the relative merits of the two models. The central place model can, of course, still be used as a rough approximation, when discussing location of facilities, even when one is not completely convinced of its universal applicability.

1.3 Criticisms of the basic assumptions

The weakest points in the central place model are undoubtedly the basic assumptions. The distribution of population depends, to a high degree, on physical factors such as topography, soil quality and availability of fresh water. The distribution of population is therefore not even. One can also assume that random variations in population density occur, even throughout a physically homogeneous district. In the same manner it is only in exceptional cases that the transport surface within a district is homogeneous. Topographical barriers such as lakes, marshes and bogs make it impossible to travel in certain directions, it is necessary to go round them instead. Rivers and streams can only be crossed at fords. A large lake may require the use of a boat, with loading and unloading of goods which requires time and energy as a consequence. Perhaps the transport surface was homogeneous at the very beginning, but a system of tracks

and simple roads soon develops. This development occurs at a very early stage, probably before the arrival of the first settlers. This means that when the central places start to develop a crude road network already exists which must naturally effect the choice of location.

If the assumption is made that the basic units of settlement lie at the nodes of a regular triangular net one must also assume that the units of settlement are located as far as possible from one another and that no contact occurs between them. This second assumption makes the growth of central places within the area in question impossible. The growth of a system of central places requires highly developed contacts between the different units of settlement. If one reckons with contacts between the different units of settlement from the very beginning, as well as a certain amount of contact with the surrounding world, it would seem likely that one should attempt to minimize the length of the track and road network within the area. Much work and time is necessary to maintain even the simplest network. If this, in fact, is the case it would seem likely that the original pattern of location of units of settlement was rectangular and not triangular. The triangular alternative results in a network that is about 25 % longer than that resulting from a rectangular pattern.

Morrill (1965) uses a triangular net in his simulation of the growth of the road network in Småland, a part of southern Sweden. The network he gives does not at first glance present a natural appearance. It would be of great interest to repeat Morrill's simulation using a square net, a rectangular net and a completely irregular net. One could then determine whether one of the resulting networks appeared to correspond with reality.

The majority of central place models assume that the umlands cover the whole district without overlapping, also that every umland for a good or a service is of the same size and shape. If both of these assumptions are correct then the resulting umlands must be regular triangles, squares or regular hexagons. Of these alternatives the hexagon is usually chosen, due to its similarity with the circle, and to the fact that the mean distance from a point within the hexagon to the centre is shorter than the corresponding distance for a square. A calculation of this mean distance for points evenly distributed over the surface of a square, with sides $2S$ long, gives as result the value $0.765S$. When the calculation is repeated for a hexagon, having the same area as the square, the mean distance is $0.755S$. The difference between these two values is only $0.01S$. This difference is small enough to make the square pattern a realistic alternative to the triangular. Application of Christaller's reasoning, using a rectangular pattern of location for the units of settlement, shows that central places might also lie according to a rectangular pattern.

2 Spontaneous location

The majority of central places in Sweden, together with most of those in western Europe, were founded by either the religious or secular authorities (Améen 1964). In the same manner Sandblad (1949) categorizes the towns of the middle ages in Skåne into king towns and bishop towns. In neither of these cases was the founding of the town and its choice of location spontaneous. The secular authorities paid close attention to strategic requirements when founding new towns. These towns were mainly for defence. Stockholm, founded in the 13th century, is a good example of this policy. The authorities also attempted to concentrate trade and crafts to the towns. A regulation forbidding the carrying on of these activities outside the towns was introduced (*näringstvång*). This was not done to encourage the growth of towns, but instead to simplify the control and taxation of trade and crafts. Even though this regulation existed a certain amount of trade and craft activities were carried on in country areas. Many parishes had their own parish craftsmen. These men often travelled between the different farms, making necessities from raw materials produced on the farms. An example of this type of craftsmen was the parish tailor who made clothes from textiles woven on the farm; another was the parish shoemaker who made shoes from the hides of slaughtered farm animals (Hanssen 1952). The competence of these parish craftsmen was not regulated in any way. Instead the villagers applied to the local court for permission to adopt a certain person as a parish craftsman.

In Sweden there existed three legal ways in which trade could be carried on in country districts. The first way was practiced by the citizens of Västergötland who enjoyed special permission to travel around between farms in all parts of Sweden selling goods. Trade was also carried on by the proprietors of industrial communities. This trade assumed quite large proportions. As a result the town tradesmen protested, but without success. Regulations published in 1642 and 1700 secured the privilege enjoyed by these communities whereby goods could be purchased in bulk and sold to persons having an association with the community (Nordström 1952). The third way in which trade could be carried on was at markets outside

the towns. The time and place for a market was determined by tradition, i.e. that a certain village had a market on the first Friday in every month. Tradesmen from the towns came to the village on the day in question and traded with the local population.

In addition to the legalized service carried on in country districts by parish craftsmen, travelling salesmen from Västergötland, and at markets and in industrial communities, there was also hidden trading activity. In most cases the practitioners were not employed full time in this activity. (Nordström 1957). A farmer that was good at slaughtering animals also slaughtered his neighbour's animals. A farmer's wife that was good at baking even baked cakes for sale. Trade with purchased goods also occurred. A number of stories from northern Skåne describe how travellers bought herring from fishermen in Blekinge and sold them on the way home, or in their home village.

A number of partial relaxations in the regulations regarding trade in country districts were made in 1846 but even still it was forbidden to open shops within 30 kilometers from a town. Only when a complete relaxation in the regulations was carried out in 1864 was it possible to establish legal service facilities in country districts. Many of the service facilities which sprang up at that time had existed illegally earlier and one can assume that their locations had come about spontaneously. Their pattern of spatial location has shown itself to be extremely stable over a long time. The Örkened district in northern Skåne is used here to illustrate such a pattern of spatial location.

2.1 Business activity in Örkened at the end of the 19th century

Even though the Örkened district was rich in forests the major industry was agriculture right up to the middle of this century. During the period when sailing ships were used the district provided a certain amount of timber to coastal towns in Blekinge, including timber for masts. Wooden products such as barrels, pumps, ladders etc. were also made in the district and exported to other parts of Skåne and Denmark where the raw materials were not as freely available. As a natural result of this activity such side activities as tar production and potash production were carried on. The manufacture of chip-baskets became known in the district during the last 2 or 3 decades of the 19th century and production of these was carried on on a large scale. Nearly every farmer started his own basket-

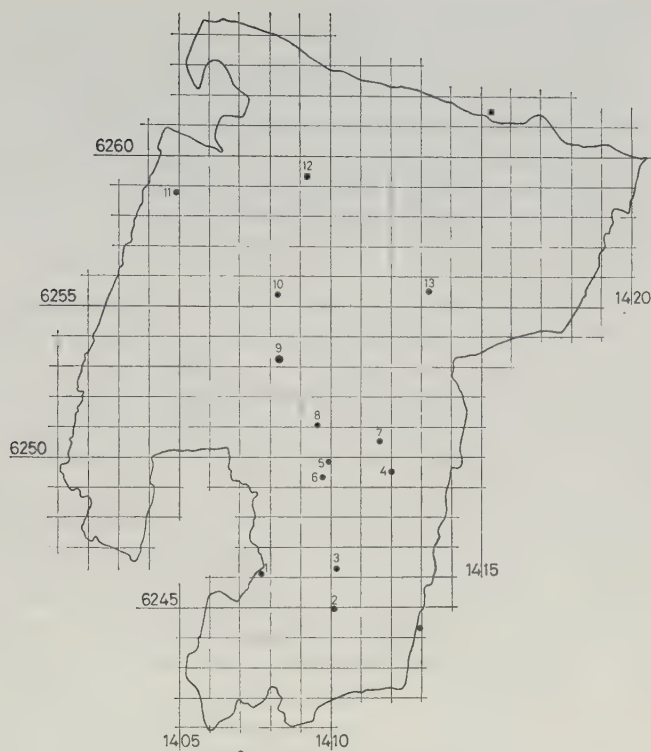


Figure 2.1. The pattern of spatial distribution exhibited by general stores in Örkened during the 1920's.

making factory which employed the people associated with the farm during the winter months. Export of the completed baskets was not limited to Skåne and Denmark, even Germany, Poland and England were common destinations. The scale of operations became so large that the manufacturers started their own export businesses. Branch factories were also opened in Denmark and Germany.

At the end of the 19th century black granite mining was begun in Örkened. The most important discovery, Gylsboda, is marked by the number 5 on the map shown in figure 2.1. The quarried stone was transported by horse and cart down to the northern end of the Immeln lake (number 2 in figure 2.1), where it was loaded onto barges and transported to the railway station at Immeln. Eventually the railway line from Sölvesborg to Älmhult was constructed and Gylsboda became a station on this line. The quarrying activity was on a very large scale and employed more than 300 men at the height of its activity.

2.2 The quarry workers' dwellings

The quarry was a large work place and one could therefore have believed that a large built-up area would grow up around it. This was not the case. The company built a number of villas for the quarry officials and two barracks for the workers, but these did not provide a dwelling for more than a small proportion. The majority of the workers were therefore forced to find a dwelling elsewhere, for example on farms in the district.

It appears that the company did not wish to become engaged in the provision of dwellings. One would have thought that they could have sold land in the vicinity of the quarry to the workers, thus making the development of a town possible. This was not done due to the fact that land might be required for waste from the quarry; also the precise extent of the find was not known.

The workers were, for their part, not particularly interested in building their own house on small sites in the vicinity of their workplace. Black granite is a typical luxury article and its production is therefore extremely sensitive to changes in the economic climate. Small scale economic crises led to reductions in the amount of granite produced, with unemployment amongst the quarry workers as a result. The workers were therefore forced to find work elsewhere. Such a move was more difficult if one owned one's own house in the vicinity of Gylsboda. The individual worker only dared to purchase his own house when he had resources for a longer period of unemployment. This type of security could be obtained by membership in an unemployment insurance scheme, but at the beginning of the 20th century this type of scheme was uncommon. However, the individual worker obtained a certain security from the site he chose to build his house on. He required that the lot be quite large with soil of good quality. The growing of potatoes, vegetables and fruit for household needs then became possible. Enough room for a cow or a pair of goats was a further requirement, providing milk and meat for the household without the need for purchasing fodder from other sources. As a result the sites were large, 2 or 3 hectares was a highly usual size.

A second demand was that the house be located near to a lake or river, thus making possible limited fishing for household needs. This possibility contributed to the chances the individual worker had of surviving a period of unemployment.

A further factor was the location of the house in relation to the place of work. The houses were all located at a fair distance from the quarry. However a travel time to work of more than 1 hour was quite acceptable, resulting in a distance of 7—8 kilometres if walking and twice the

distance if cycling. Nearness to the workplace was rated low. The same phenomenon can be studied in existing towns. An investigation in Malmö into the reasons behind an individual's choice of dwelling showed that only 5 % of the sample chose with regard to nearness to their work place (Lindberg 1971). Of those wishing to move from their present dwelling only 2 % desired to do so due to the fact that their present dwelling was too far away from their work place.

2.3 Service facilities in Örkened 1920 and 1970

As result of the removal of the restrictions on trade in country districts a spontaneous location of such service facilities as general stores, tailors and shoe makers occurred in those districts. According to the central place model these facilities should be located as far as possible from one another and assume positions at the nodes of a triangular network. Figure 2.1 is a map showing the distribution of general stores in Örkened about 1920. As can be seen the spatial pattern they describe is not remotely a regular triangular net. The data for this map has been obtained from interviews with emigrants from and immigrants into Örkened about 1920. The spatial pattern shown in figure 2.1 was very stable and was not altered to any degree before the 1950's.

Figure 2.2 is a map showing the distribution of population in Örkened in 1920. There was a total of 4714 persons registered there on the 1st of November that year. The population had decreased by about 10 % compared with just before the first world war. This decrease depended a great deal on the fact that quarrying of black granite had stagnated during the war.

The source for the map shown in figure 2.2 was the "mantalslängder" (population registration lists) for Örkened for 1920. These lists were used to determine the population of the different hamlets and villages, see table 2.1. In the next stage a square grid net with squares 2.5 km by 2.5 km was placed over the area and a decision was made regarding within which squares the different villages lay. The population has then been added square for square to produce the map shown in figure 2.2. Unfortunately the squares are rather small in proportion to the villages. This has resulted in some squares with no population at all, as the whole population of each village has been placed in the same square. If it had been possible to use real estate data the number of empty squares would have been reduced. This would also have resulted in a reduced population in those squares now having a somewhat too high population. Square (1407.5, 6245.0) which now contains 411 persons living in the villages of Strönhult, Råge-

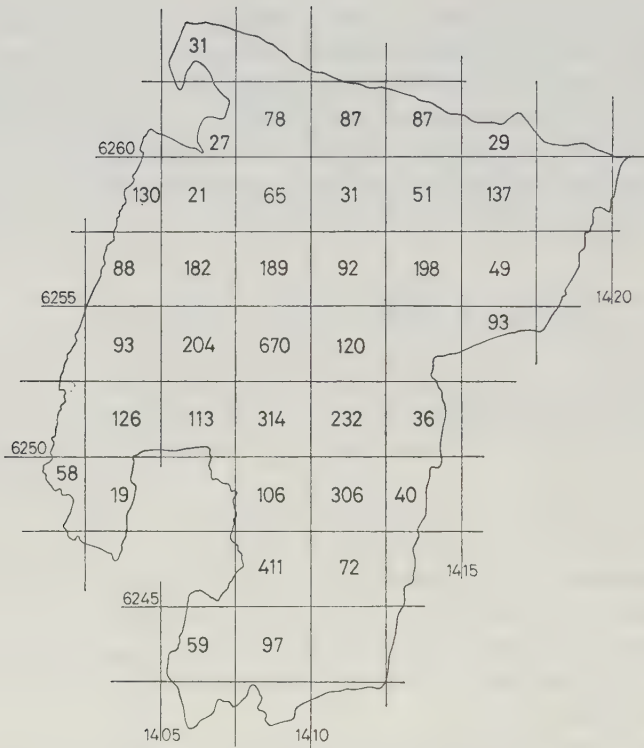


Figure 2.2. The distribution of population in Örkened in 1920. Square size: 6.25 km².

boda, Kruseboda and Grimsboda should in fact have its population divided up between 4 or 5 different squares. Even with regard paid to these inaccuracies it is possible to see that the southern part of the local authority district is more densely populated than the northern part.

The central place in Örkened is Lönsboda (number 9 in figure 2.1). In 1920 this town had a population of 300 persons. The service provision was completely out of proportion to the population. A list of the shops in this small town includes 3 clothing shops, 5 general stores, 3 butcher's shops, 1 fish shop, 1 paper, dye and drug store, 1 ironmonger's, 1 wool shop and 1 watchmaker's shop. In addition there were 1 bank, 1 hotel and 7 cafés, see table 2.2. Several of the cafés were run by widows or elderly unmarried ladies who had moved in from the country. One can look upon these cafés as a form of pension which helped to supplement the income from capital inherited from their husbands or parents. The café was usually located in the owner's dwelling and the dwelling was therefore used for two purposes. Another type of service facility which

	1920	1970		1920	1970
Applehult	58	3	Kvarnatorp	24	19
Björkhult	97	37	Kärraboda	71	26
Björstorp	19	3	Kätteboda	49	8
Bosakulla	35	5	Lönsboda	380	2074
Böglarehult	122	61	Nedaryd	39	6
Bökön L	18	10	Nybygden	19	1
Bökön S	54	28	Nyteboda	45	5
Duvhult N	68	20	Olastorp	24	19
Duvhult S	56	13	Rumpeboda	58	12
Edema	113	41	Rågeboda	103	32
Ekeshult	42	6	Rävatorp	21	11
Ekhult	58	6	Rönneboda	34	15
Ekön	24	5	Santatorp	36	5
Esseboda N	34	29	Skinnellyra	60	23
Esseboda S	66	14	Skyttetorp	57	22
Flyboda V	129	159	Smålatorp	22	6
Flyboda Ö	161	86	Strönhult	206	80
Frostentorp	27	0	Svanshult	35	23
Fröatorp	40	12	Tjuvön	79	10
Gisslaboda	314	131	Tommaboda	160	71
Graveboda	51	9	Tommahult	44	34
Grimsboda	46	17	Tostaboda	82	35
Gylleboda	3	0	Tostaholm	13	3
Gylsboda	149	63	Tosthult	198	71
Havhult N	33	7	Traneboda	106	43
Havhult S	46	16	Tranetorp	35	8
Hässlehult	8	0	Trulsatorp	29	25
Hunshult	120	59	Trånghylltan	9	4
Hägghult	92	71	Ubbaboda	150	58
Hövidstorp	8	1	Ulvshult	93	22
Johnstorp	27	11	Vermanshult	49	19
Komålen	28	13	Åbuen	133	81
Kruseboda	56	17	Örnanäs	17	2
Kungsgården	59	11	Summa	4741	3837

Table 2.1. The population of the different hamlets and villages in the Örkened area in the years 1920 and 1970. The population of Lönsboda built-up area was about 300 persons in 1920 and 2269 in 1970. Source: Population registration lists for the Örkened area for 1920 and 1970.

used the dwelling for two purposes was that referred to as "part time service" by Nordström (1965). In this case the practitioner had another job but used his free time for the selling of such things as insurance or farm machinery. Usually a very limited range of goods was carried often in highly unusual combinations. An example of this is provided by the person that sold seeds and underpants and nothing else.

Post office and railway station were also available together with other service facilities of religious and secular art. There was at this stage no doctor or dentist, although a nurse was stationed in Gylsboda.

During the period 1920—1970 the population of Örkened decreased even further. As can be seen from table 2.1 and figure 2.3 the population of

	1920's	1970's
Clothing shop	3	3
General store	5	3
Butcher's shop	3	—
Fish shop	1	—
Greengrocer's shop	—	1
Dye and drug stores	1	2
Paper and tobacco stores	1	2
Ironmonger's shop	1	1
Wool shop	1	—
Watchmaker's shop	1	1
Furniture shop	—	2
Radio shop	—	2
Electrical shop	—	1
Bank	1	3
Hotel	1	1
Café	7	1

Table 2.2. The number of different stores and shops in Lönsboda in 1920 and 1970.

the area decreased to 3837 in 1970. A reduction of about 20 % compared with 1920. This occurred at the same time as Lönsboda increased its population considerably. The result was a reduction in the number of residents in the country districts, this reduction showed itself in the form of a thinning out rather than total disappearance of population from particular areas. The small general stores remained suprisingly long. The last of them were not closed until the 1960's when they were replaced, or driven out of business by mobile shops. At the moment the population of Örkened is at a sufficiently high density to allow mobile shops to be profitable. A further reduction in the population, or a rise in the cost of running mobile shops, even of quite a small size, will certainly result in the number of customers being insufficient for this method of distribution.

The railway from Sölvesborg to Älmhult carries traffic at the moment but will probably be closed down fairly soon. Lönsboda is the only railway station in the area whereas in 1920 there were 5 stations. Reductions have been the same in the case of post offices. In 1970 the only one left was in Lönsboda. The other post offices have been replaced by country postmen. However, from an overall point of view, the population of Örkened has a better level of service now than in the 1920's even though shops and post offices have been closed.

Lönsboda now has 1 health clinic, 1 dental clinic, 1 private dentist and 1 chemist's. The health facilities available within the district have improved considerably since the 1920's.

The majority of the part time service has disappeared between 1920 and 1970. The number of shops in the town has also decreased during this

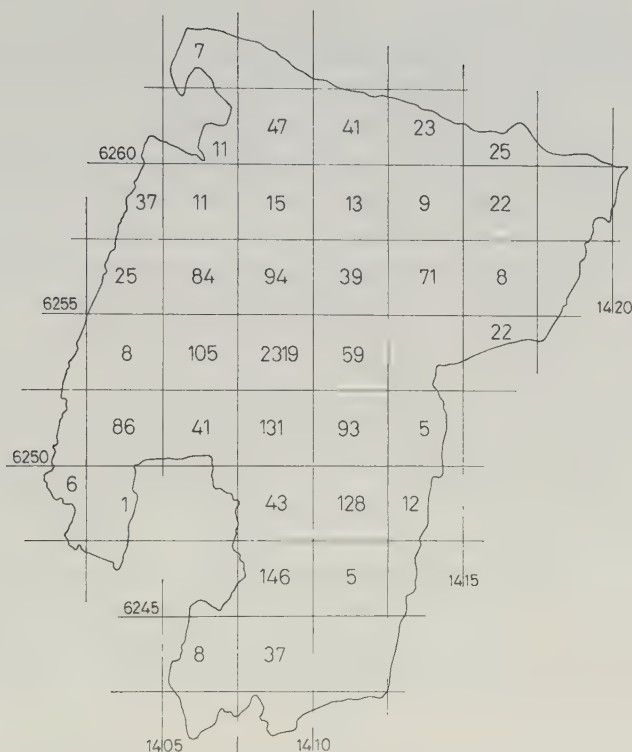


Figure 2.3. The distribution of population in Örkened in 1970. Square size: 6.25 km².

period. See table 2.2. There is now only 1 café instead of the 7 existing in 1920. Completely new types of shops have been opened such as 2 furniture shops, 2 radio shops, 1 electrical shop and 1 greengrocer's shop. 3 fairly large general shops, supermarkets, sell all sorts of food, meat and fish included. Lönsboda in 1970 also boasted 3 banks, 1 hotel, 1 ironmonger's, 2 dye stores, 2 paper and tobacco shops, 3 clothing shops and 1 watchmaker's shop. From an overall point of view this little town provides a wide range of goods and services. It could be considered ideal, seen from the consumer's point of view, with its wide range of services within a reasonable distance from his dwelling.

In 1920 there existed a number of tailors and shoemakers in the country districts of Örkened. A map of their distribution shows a pattern similar to that for general stores. The number of tailors, shoemakers and clog makers is less than the number of general stores and as a result the pattern is more sparse. One can also observe that no combined location of these occurs outside Lönsboda.

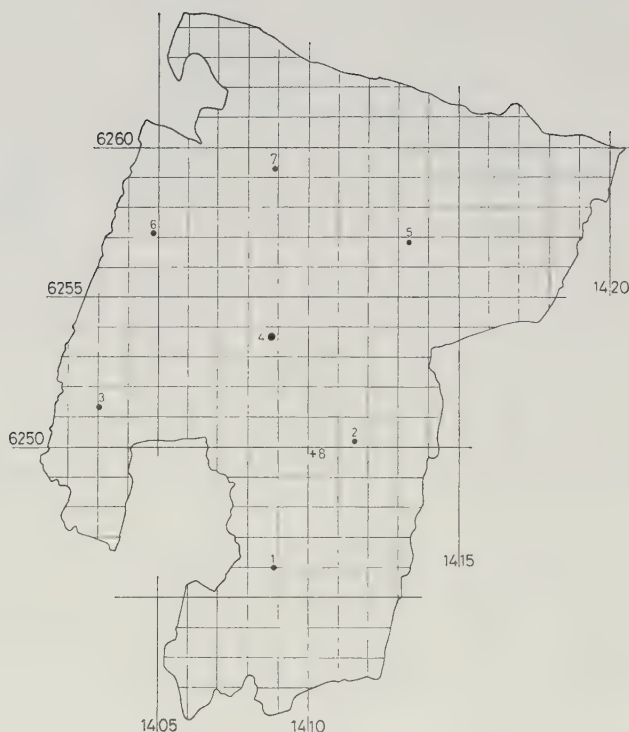


Figure 2.4. The pattern of spatial distribution exhibited by schools in Örkened during the 1920's. The school in Gylsboda (8) contained only primary classes (classes 1 and 2) and functioned only for a short period. The remaining schools had the full 6 classes.

Amongst those service facilities which are not spontaneously located can be included schools. In order to build a school a decision must be made by some authority, a parish meeting for example. As can be seen from figure 2.4 these decisions resulted in a pattern of school locations which is very near to the theoretical triangular network. The impression that one obtains a nearly perfect triangular net is accentuated when the schools in the districts abutting on to Örkened are included. Rationalization has also affected the schools and the village schools are now closed. The children all go to the central school in Lönsboda.

2.4 The development of central places and the spontaneous location of service facilities

Other reasons than spontaneous location appear to lie behind the location and the growth of central places. Originally there were only a few work-

places in Lönsboda, but it had a church and a certain amount of religious and secular service facilities. These functions probably lie behind Lönsboda's development into a central place. Other investigations also show that villages having churches often developed into more and more all-round central places while other villages in the vicinity stagnated (Nordström 1957). The decisions made by religious and secular authorities concerning the location of churches, schools etc. and the requirements these authorities had for ease of communication with the general public, probably provides more of an explanation of the pattern of central places than that given by spontaneous location. Against this must be set the fact that Örkened is perhaps too small an area to be used for drawing far-going conclusions regarding the relationship between the central place pattern and spontaneous as against planned choice of location. It can be said that the spontaneous location of general stores in Örkened gives a totally irregular pattern, but that the planned location of schools in the same district gives a nearly perfect triangular pattern. If this empirical rule is general, i.e. if it applies to all areas and at all levels, then central places should be located in a more regular triangular pattern the less spontaneous the forces behind the choice of location is. In any case it would appear worthwhile to investigate the pattern of location of Swedish built-up areas in order to determine whether the pattern is in some way regular or irregular.

3 The central place pattern in southern Sweden

A prerequisite for any study of the pattern of location of built-up areas is data describing the distribution of population within the area. A built-up area can be defined as an area with a high population density when compared with its surroundings. The boundary between a built-up area and a country district goes at the point where a noticeable increase in population density occurs (Nordbeck 1969). At every central place a population peak occurs, the central place theory states that these peaks recur periodically. An analysis of the spatial pattern of central place location is best begun by seeing whether the population density varies periodically or not. In the latter case the central place pattern is irregular. If, on the other hand, it has a regular form then one proceeds to determine the period. This information can then be used to determine the pattern itself i.e. if the central places lie at the nodes of a triangular, square or rectangular net.

3.1 The distribution of population in southern Sweden 1965

The basis for an analysis of the central place pattern is an accurate map of population distribution in the area to be studied. Southern Sweden is therefore a particularly good area for investigation, as square net maps of population distribution for the year 1965 are available for the counties of Malmöhus, Kristianstad, Blekinge, Kronoberg and Halland. The grid squares on all of these maps are 1 km². The distribution of population is therefore given sufficiently accurately to make possible an analysis of the central place pattern.

Of these 5 maps, the maps of the counties of Halland, Kronoberg, and Blekinge are based on coordinate based real estate data. The coordinates used have an accuracy of 1 km. In the case of Kristianstad and Malmöhus counties parish data was used, but the population living in the country districts of every parish was distributed spatially in accordance with the

distribution of houses in these areas. The distribution of houses was obtained from the topographical map of Sweden. This series of maps was started in the middle of the 1950's. Those sheets covering Kristianstad and Malmöhus counties were published in the years immediately preceeding 1965. The maps therefore give an accurate picture of the distribution of buildings in the two counties making up Skåne about the middle of the 1960's. This means that it is probably not possible to construct a more accurate map of the distribution of population without access to coordinate based real estate data.

3.2 The size of the reference area and the equidistance of the grid net

A reference area of some size is always necessary when calculating population density at a point. If a square is used as the reference area and no overlapping is allowed a square net map is obtained. The reference areas can be organized in a rectangular or triangular grid net (Nordbeck 1964b). In the second case every alternate row is displaced half the equidistance of the grid net. As the input data is given in the form of a square net map it seems most suitable to use a reference square in the analysis of the central place pattern. A number of squares, for example 4, 9, 16, 25, can then be combined to form larger squares.

An analysis of the pattern of distribution of central places is highly dependent on the size of the reference area chosen. If the length of side of the reference area is equal in size to the period one is searching for, the period will not show in the results. The reference area should therefore be considerably less than the period. According to Christaller (1933) the distance between two nearby hamlets is about 7 km. If one wishes to check this hypothesis the size of the reference square should not be more than half of the distance, i.e. about 3 km. The equidistance of the grid net was fixed at 2 km which made possible the use of a regular triangular grid net. As the size of the reference square should be a multiple of the equidistance the 3 km reference square cannot be used. In this case 2 km was used with 4 km as an alternative.

Christaller (1933) also states that the equidistance occurring in connection with the pattern of distribution of township centres is 12 km. The analysis of this pattern therefore should use a reference square with sides 6 km long. The next level in the central place hierarchy is the county seat and the equidistance given by Christaller in this case is 21 km. A suitable size for the reference square would be 10 km in this case. A further stage up

the hierarchy brings one to the district city with an equidistance of 36 km, so a suitable reference square in this case is 18 km.

The size of the reference area has been consistently chosen to be less than or equal to half the equidistance of the respective pattern. One, of course, uses the technique with overlapping reference areas when calculating the population density at different grid points. As the equidistance of the grid net is always 2 km the overlapping is equal to 1, 3, 5 and 9. A multiplication of the overlapping with the grid equidistance gives the length of side of the reference square.

3.3 The testing of the central place pattern

Probably the best known method of investigating periodic features in a data series is the auto-correlation method. In this method the original series is duplicated and the duplicate displaced one step relative to the original. The first element in the new series corresponds with the second element in the original and so on. The correlation coefficient of the two series is then calculated. The duplicate series is then displaced a further step and the correlation calculated once more. This procedure is repeated until so few values remain in the duplicate series that a calculation of correlation is not worthwhile. If the given series is periodic the correlation coefficients will also be periodic. A displacement of the data a whole period gives correlation coefficients near to +1, coefficients near to -1 indicate that displacement is about half a period. Auto-correlation has been mainly used for the analysis of time series but can also be used for the analysis of spatial data. Auto-correlation techniques are normally very time-consuming and it is therefore necessary to use a computer. Several auto-correlation programs exist, amongst them can be named those included in the BMD package (Dixon 1968) and NORAUT, which is a program in the NORMAP system. The latter program is designed to be used in the analysis of spatial data.

Auto-correlation analysis requires the use of a fairly large study area, but on the other hand this area should not be too large. Too large an area leads to exceedingly long calculation times. It was considered that a population square net map containing 160 rows and 160 columns was best suited to the analysis of periods in a population distribution. The squares in the grid are 1 km² each which gives a map covering the whole of Skåne, southern Halland, southwestern Småland and western Blekinge. The actual auto-correlation is performed in an area 64 columns by 64 rows. The equidistance of the grid used is 2 km which results in the auto-correlation

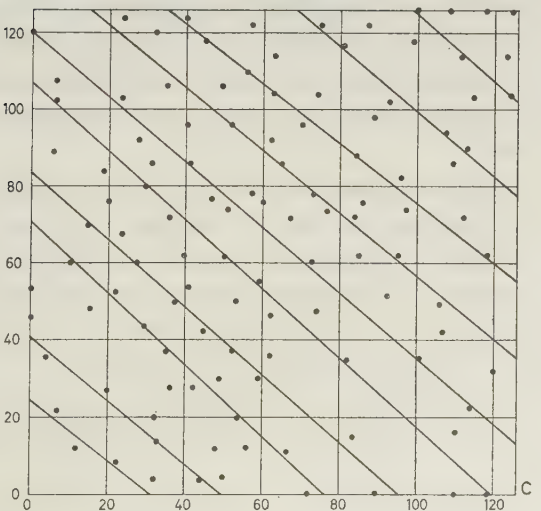
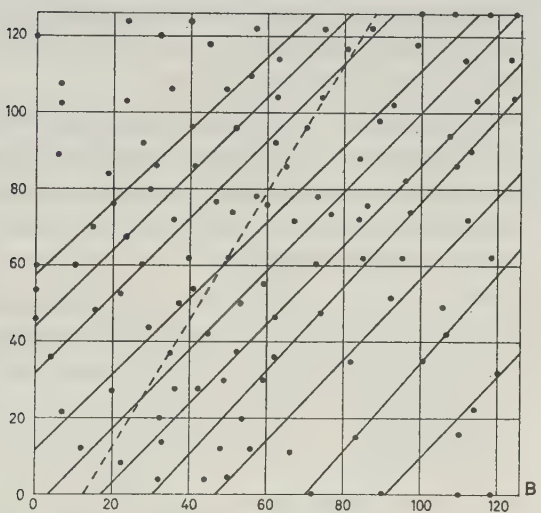
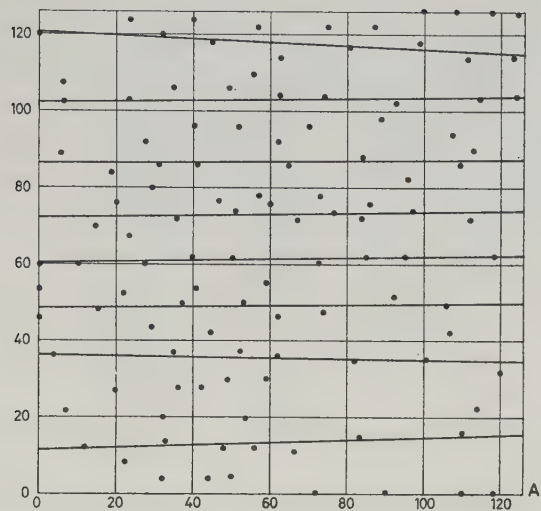


Figure 3.1. The pattern exhibited by the maximum points of the auto-correlation function, when applied to southern Sweden. The reference area is a square with sides 2 km long. Data: Square net maps of population distribution in Skåne, together with Blekinge, Halland and Kronoberg counties.

	A	B	C
1	2°	46°	142°
2	-1°	49°	141°
3	1°	47°	137°
4	1°	47°	139°
5	1°	46°	138°
6	0°	46°	140°
7	1°	45°	141°
8	-2°	45°	142°
9		45°	140°
10		44°	139°
mean	0°	46°	140°

Table 3.1. The angles between the x axis and the lines shown in figure 3.1.

analysis being performed for a square with sides 128 km. As was mentioned in section 3.2 the analysis was performed for the 4 sizes of reference square 4 km², 36 km², 100 km² and 324 km². In all these cases the equi-distance of the grid net is 2 km.

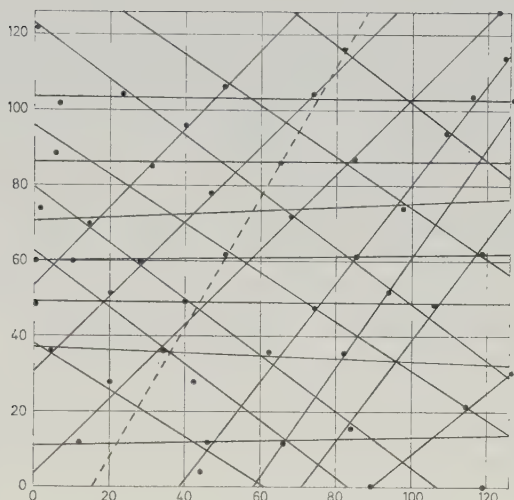
In the first case, where the squares are 4 km², the auto-correlation coefficients are relatively low. Only in two cases are they as high as 0.5. This indicates that the data is not of a particularly periodic nature or that its periodic nature is so complex as to prohibit its discovery with simple auto-correlation analysis.

As was pointed out earlier the correlation coefficients should vary between -1 and +1. In the example studied here the range is considerably less. Even though the range is small certain maximum points can be located. A maximum point is, in this case, a point at which the correlation coefficient is higher than at all the points surrounding it. As a result of the fact that a large number of points were used when calculating the correlations (200—6400) the maximum values obtained are significantly more than 0 even though they are quite low. The maximum points obtained are illustrated in figure 3.1. The x axis of this diagram refers to the amount of displacement in the east—west direction in kilometres. In the same way the displacement in a north—south direction has been shown on the y axis.

Auto-correlation analysis also involves study of the pattern described by the maximum points. If these points lie at the nodes of a regular triangular net then it should be possible to enter three sets of parallel lines into diagram 3.1. Every line should go through, or pass in the vicinity of, a number of maximum points. Also the angles between the lines in the different systems should be multiples of 60°.

A study of figure 3.1 shows that such lines pass in the vicinity of the lines $y = 36$, $y = 60$, $y = 86$ and $y = 104$. This system is referred to as

Figure 3.2. The pattern exhibited by the maximum points of the auto-correlation function. Reference area 36 km².



system A. Figure 3.1 A shows 8 lines which form part of this system. Table 3.1 contains the angle between each of these lines and the x axis. The lines in system A lie nearly parallel with this axis. The next stage is to attempt to place a new system of lines in the diagram. This system, system B, should lie at an angle of 60° to the lines forming system A. A line of this kind can be detected in the upper part of the diagram (the dotted line in figure 3.1). It lies very near to the route taken by the railway from Malmö to Stockholm. If the requirement that all the lines in system B should pass through the maximum points is to be satisfied a much smaller angle between system A and B than the expected 60° must be used. This version of system B is illustrated in figure 3.1 B and the corresponding angles are given in table 3.1. In the same way a third system, system C, is added to the diagram. As can be seen from table 3.1 the angle between A and C is about 140° . The angle between B and C is 94° , considerably more than

	A	B	C
1	0°	39°	142°
2	0°	53°	145°
3	1°	55°	143°
4	1°	52°	147°
5	0°	45°	143°
6	-2°	44°	143°
7	-2°	46°	148°
8	2°		
mean	0°	48°	144°

Table 3.2. The angles between the x axis and the lines shown in figure 3.2.

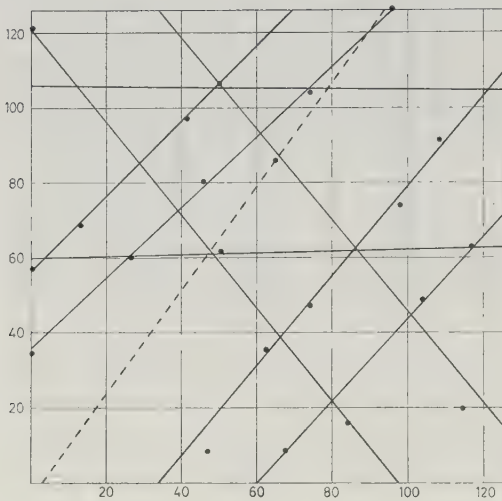


Figure 3.3. The pattern exhibited by the maximum points of the auto-correlation function. Reference area 100 km².

the 60° one would expect from the central place model. Also a further system, system D, could be added to the diagram at 90° to system A. An auto-correlation analysis of population density in southern Sweden supports the view that the pattern of location of central places in the area is therefore just as likely to be rectangular as it is to be triangular.

An increase in the reference area's size to 36 km² causes a considerable reduction in the number of maximum points obtained from the auto-correlation analysis. Figure 3.2 shows that three systems of parallel lines can still be constructed. The angle between systems A and B is about 48°

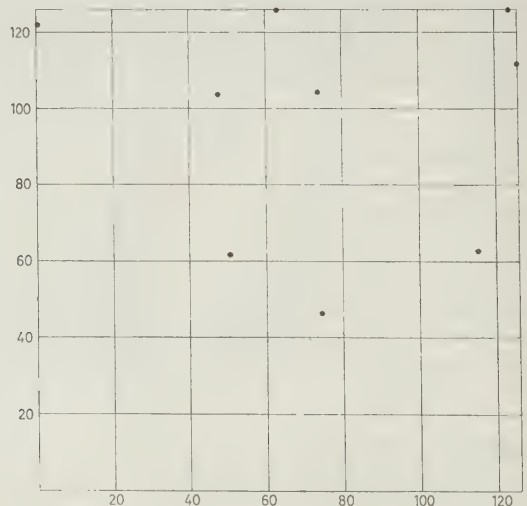


Figure 3.4. The pattern exhibited by the maximum points of the auto-correlation function. Reference area 324 km².

and that between B and C is about 96° , see table 3.2. Also in this case the pattern obtained appears more rectangular than triangular. The only exception to this pattern is the dotted line, i.e. the railway line from Malmö to Stockholm. This line does not appear to fit in the lower half of the figure.

The number of maximum points is reduced even further when the size of the reference square is increased to 100 km^2 , see figure 3.3. The number of maximum points is now so small as to make it necessary to restrict the number of lines, compared with figures 3.1 and 3.2. The tendency is still the same as that shown in the two previous figures. Even at this level the pattern of central place location can just as well be rectangular as triangular. Even in this case the railway line between Malmö and Stockholm deviates from this pattern.

An increase in the length of side of the reference square to 18 km (the district city level in the hierarchy) reduces the number of maximum points even further, as is illustrated in figure 3.4. Only 5 such points lie within the area studied. This is, of course, far too few to make possible the construction of system of parallel lines. A visual inspection gives the impression that the points obtained can hardly be nodes in a triangular or rectangular net.

One can summarize this analysis of the pattern of location of central places in southern Sweden using auto-correlation by saying that the results it gives hardly indicate that the pattern is regular and triangular. If the pattern is in any way regular then it is more likely to be square or rectangular than triangular. It has therefore not been possible to verify the central place location model using data from southern Sweden. The central place model is probably too crude to be directly used as a basis for location discussions. Hence, the solution of location problems must involve the use of other models developed specially for this purpose.

4 The distribution of population in Sweden in the years 1855, 1917 and 1965

The possibility of producing maps of population in Sweden for at least 3 points in time, i.e. the years 1855, 1917 and 1965, has already been shown. In all these cases the construction of the maps has been done independently. Unfortunately the majority of Swedish population maps are not square net maps, but it is fairly easy to transfer the information they contain to a map of this type. How this transfer is carried out is described in the following section at the same time as the individual maps are described.

Square net maps can be used as input data to location models (Nordbeck & Rystedt 1971). This is a strong reason for transferring the population maps concerned to square net maps, the resulting square net maps can be used for location calculations at the national and regional levels. The maps, available for three different points in time, make possible an investigation of how changes in population distribution between 1855 and 1965 affect the solutions to location problems given by algorithms for multiple locations.

4.1 The railway investigation committee's map

In 1858 the Swedish government set up a committee to investigate the most suitable routes for the Swedish railways. In order to do this it was necessary to get information about the population distribution. This was done by the construction of a square net map of population distribution in southern and central Sweden. First a basic map with a grid network was constructed. The grid equidistance was 1 old Swedish mile, giving squares with an area of about 114.3 km². The population registration lists were then consulted and the population of every village calculated. The location of each village on the map was then found and the popula-

tion of each square in the map obtained by addition. This sum was then entered into the square in question. In order to make the map more easily readable the different squares were given colours. Different colours were used for every thousand, i.e. each of the classes, 0—1000, 1000—2000, and so on had its corresponding colour.

The railway investigation committee's original map with the population entered in every square has not been found (Nordbeck 1964c). A simplified version of the map was printed, without the population totals entered into every square, and with only the different colours shown. This printed map used a more detail division into classes than the original map. The 4 lowest classes were divided into 2, giving them a range of 500 persons, the remaining 3 classes, 4000—5000, 5000—6000, and more than 6000 were retained as before giving a total of 11 classes. The population of towns was not included in the population, instead it was given in table form at the side of the map.

The railway investigation committee's map showing the distribution of population in 1855 needed to be transferred to a new square net map with another grid net, of a different size. It was thought most suitable to use the land use map's coordinate system as this system is widely used, for example in connection with the coordinate based real estate register. The grid size chosen was 100 km². The first stage involved the construction of a basic map with this coordinate system. The construction of the new population map was done by photographing the original map on to a colour slide, which was then projected on to the basic map. The position of the projector was chosen so that the projected map coincided with the basic one as far as was practically possible.

As was described earlier every colour in the original map refers to a population density. The average population density of every square in the basic map could then be estimated from this colour information. If a square in the new map under construction had only one colour then that square could be assigned the corresponding population density from the map of 1855. In most cases the squares of the new map were made up of parts with differing colours. The proportion of each square taken up by each colour was estimated and recorded and the estimated population density calculated from this information and the class divisions of the original map. Figure 4.1 shows a square net map of population distribution in Sweden for the year 1855, the central and southern areas of this map where constructed using the above method.

An estimate of the amount of error in the new map was obtained by adding the populations of squares falling within the same county. As the population of every county is known exactly the two figures can be

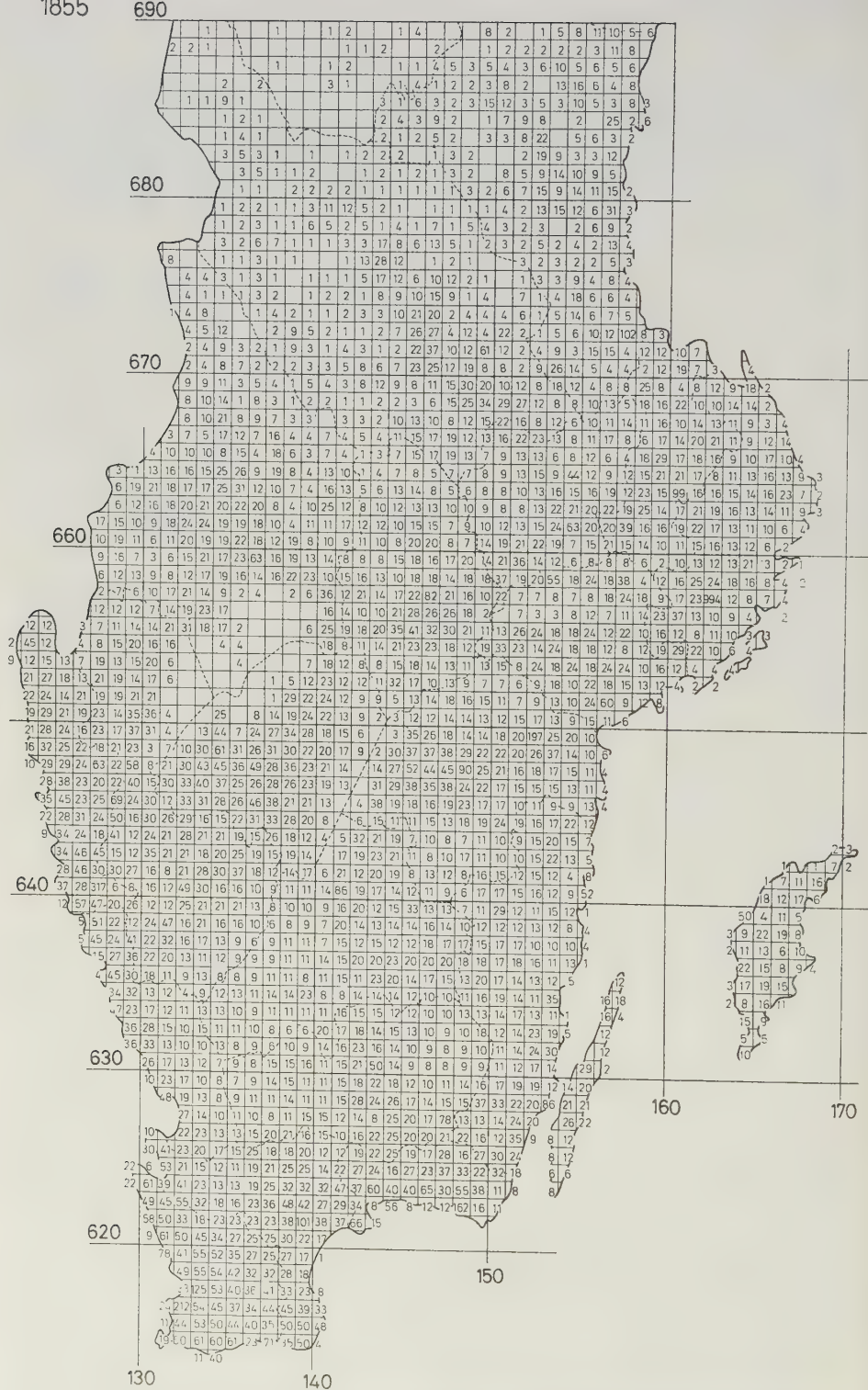


Figure 4.1. Square net map of population distribution in southern and central Sweden in 1855. The population is given in hundreds per 100 km².

1855

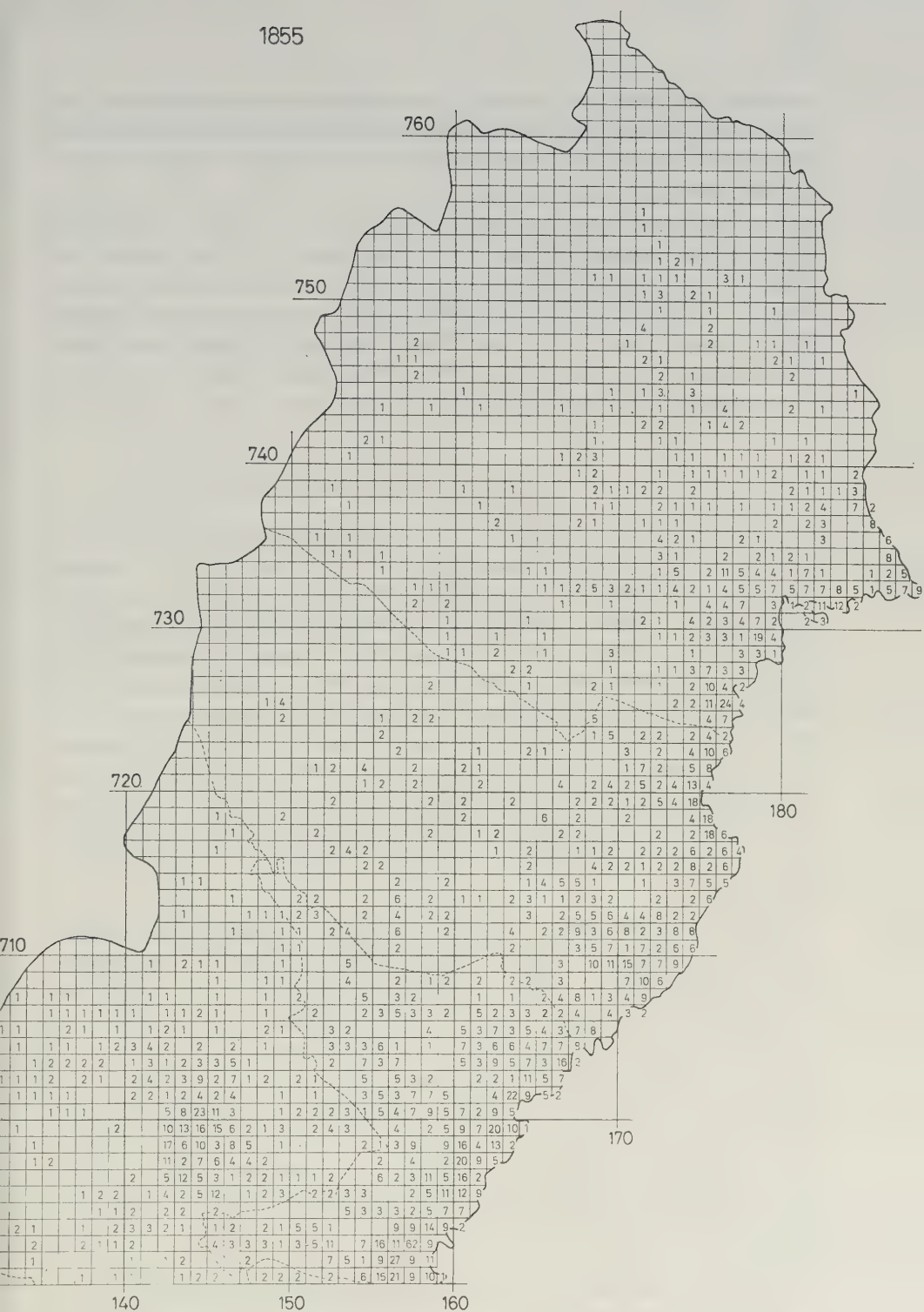


Figure 4.2. Square net map of population distribution in northern Sweden in 1855. The population is given in hundreds per 100 km².

compared. The maximum deviation between the correct population of each county and the estimated population was about 4 %. In most cases considerably less. The new map was then adjusted to give the correct population in every county.

In conclusion one can say that the 1855 map is of very high accuracy. This accuracy was obtained by the use of very small units, the hamlets and villages. Also the population of the villages was obtained from the registration lists, which are highly accurate. Only in the 1960's was this type of map prepared again for certain Swedish counties (Claeson 1964, Olsson 1966, Claeson & Stanisic 1969). None of these maps cover as large a part of Sweden as the map of 1855.

4.2 Thure von Mentzer's population dot map

The square net map of 1855 covers only the central and southern parts of Sweden. That part of the country lying north of the line $Y = 675$ needed another source from which to construct a square net map. A population dot map of the whole of Sweden and Norway is available (Mentzer 1859). This map is contemporary with the railway investigation committee's map (Kant 1970). Thure von Mentzer was a skilled cartographer and soldier and this position gave him access to the secret maps prepared by the military authorities in Sweden during the first half of the 19th century. These maps could be used to obtain a good knowledge of the distribution of built-up areas and dwelling houses throughout Sweden. He also had access to population data for the individual parishes. These two sources of information were used by him to construct a relatively accurate population dot map.

A number of difficulties are encountered when transferring his dot map to a square net map due to the fact that different dot values are used in different parts of the country. In areas with a small population density each dot can refer to less than 100 persons. In more heavily populated areas this dot value can be greater than 200 persons. An investigation of parishes in different counties showed that the same dot value is used throughout a county. Before the dot map could be transferred to a new map the value of a dot in each county was decided. The dot map was then projected on to the basic map in the same way as for the railway investigation committee's map and the number of dots per square counted. The resulting map forms the northern part of the map in figure 4.1 and the whole map in figure 4.2.

One has reason to believe that the dot map is not as accurate as the square net map of 1855. The different data and the construction method used

make this assumption reasonable. The 1855 dot map also covers central and southern Sweden and could therefore have been used as the basis for the new square net map even in this part of the country. In order to obtain a comparison between the quality of the two maps a square net map of population in certain areas of southern Sweden was prepared on the basis of the dot map. The two maps were then compared and found not to be identical, but the differences between them usually consisted of too many people in a square and too few in the neighbouring squares. The differences were thus a matter of detail and could be accepted. The two maps shown in figures 4.1 and 4.2 can therefore be used as a basis for regional analysis in those cases where the regions used are larger than 100 km².

4.3 Sten de Geer's population dot map

The without doubt best known of the Swedish population dot maps is Sten de Geer's from 1917. It is constructed in roughly the same way as Thure von Mentzer's. This means that population data for each parish is distributed throughout the area of the parish according to the distribution of buildings. Sten de Geer obtained his information regarding the distribution of buildings from topographical maps, most of them dating from the middle of the 19th century. A long time had therefore elapsed between the construction of these topographical maps and Sten de Geer's application of them in the construction of his population dot map.

Sten de Geer's map differs from the 1855 dot map in three ways. First, the scale is much larger and as a result the map covers 12 sheets. Second, the population is represented using unit dots or more accurately unit spheres. Such a sphere refers to 100 persons, where on the map it might be located. Thirdly, the population of built-up areas is shown by spheres. The volume of such a sphere is proportional to the area's population. Thus, for example, a built-up area with 100 000 population is represented by a sphere the diameter of which is 10 times that of the unit sphere. This sphere is drawn on the map at the position occupied by the built-up area.

The transformation from the 1917 dot map to the square net maps shown in figures 4.3 and 4.4 was carried out in a different way to that used in constructing figure 4.2. The coordinates of a number of points on every sheet were known. These coordinates were given in the coordinate system used for the land use map. The next involved the placing of a sheet of the dot map on to the board of an automatic coordinate measuring device (coordinatograph). The coordinates of the points mentioned previously were then determined in the local coordinate system used by the coordinatograph. These two sets of coordinates were used to determine transforma-

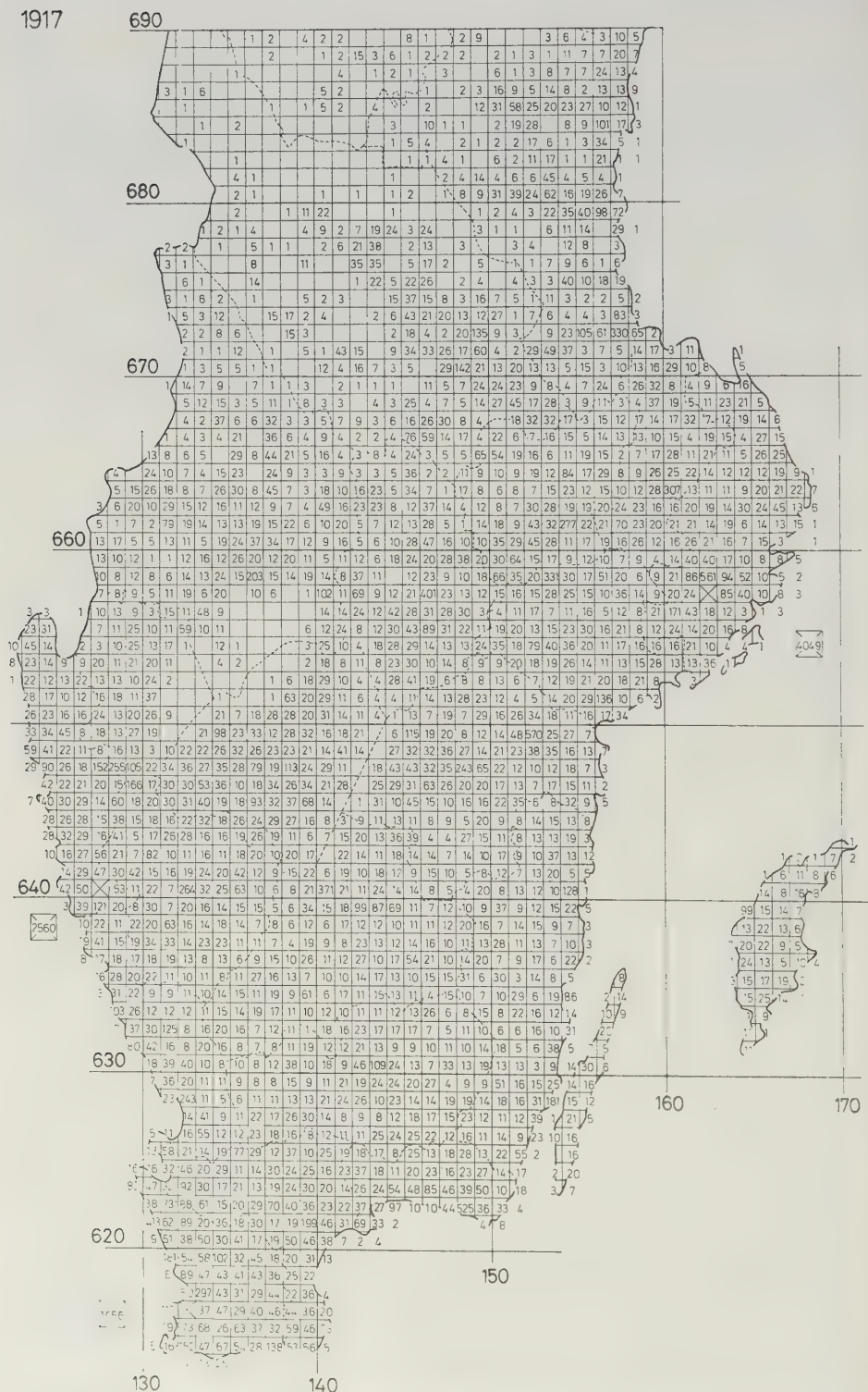


Figure 4.3. Square net map of population distribution in southern and central Sweden in 1917. The population is given in hundreds per 100 km².

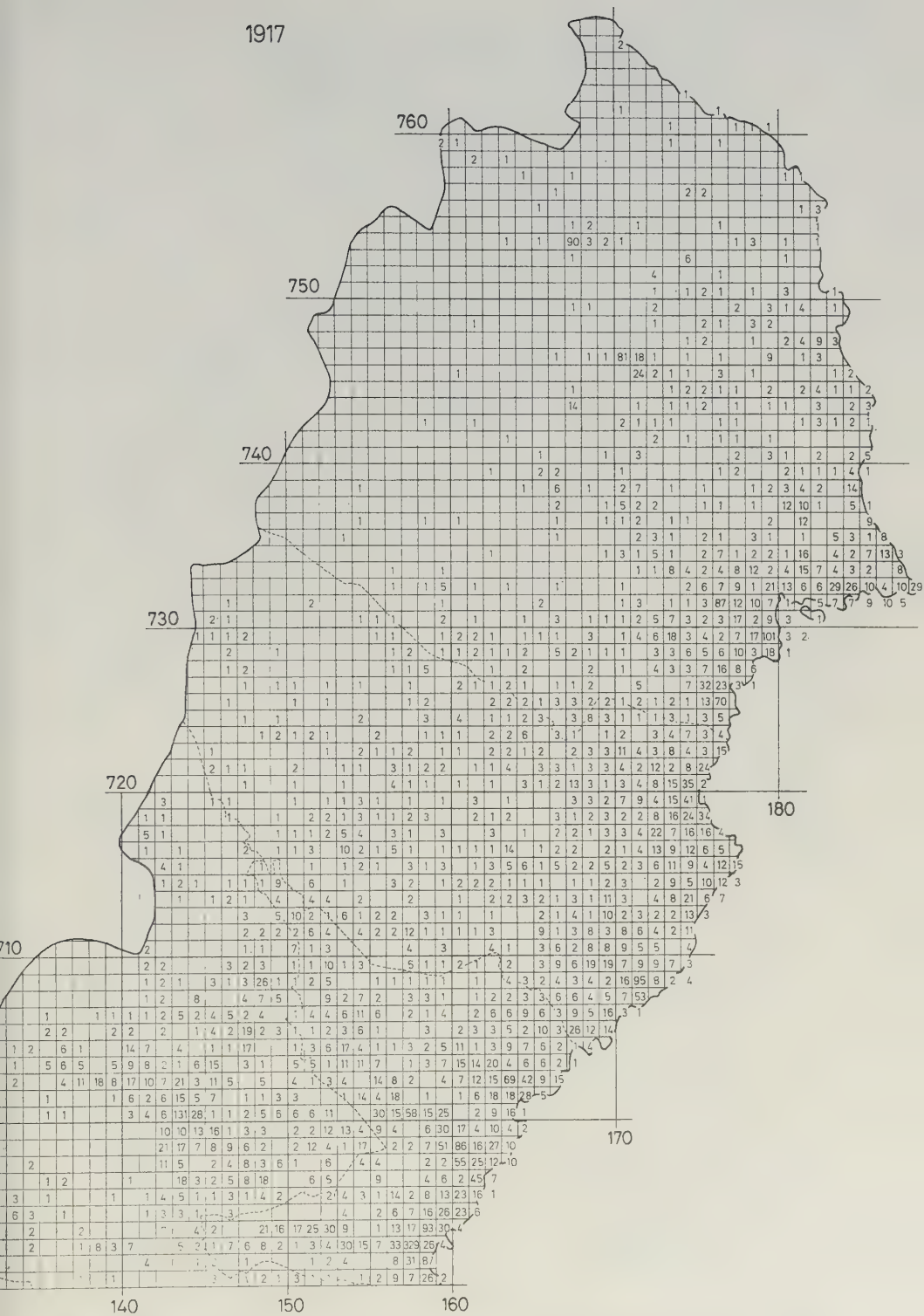


Figure 4.4. Square net map of population distribution in northern Sweden in 1917. The population is given in hundreds per 100 km².

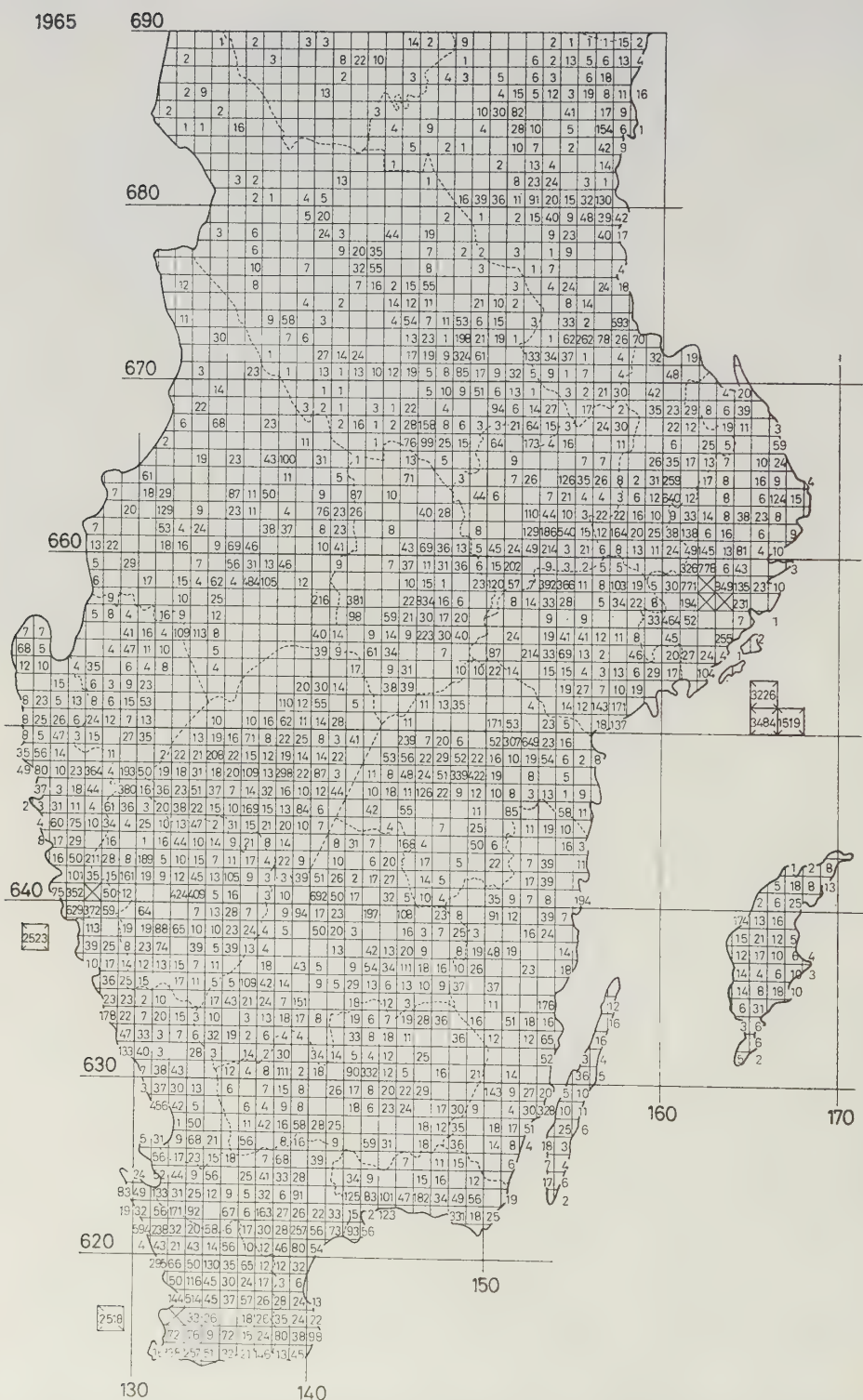


Figure 4.5. Square net map of population distribution in southern and central Sweden in 1965. The population is given in hundreds per 100 km².

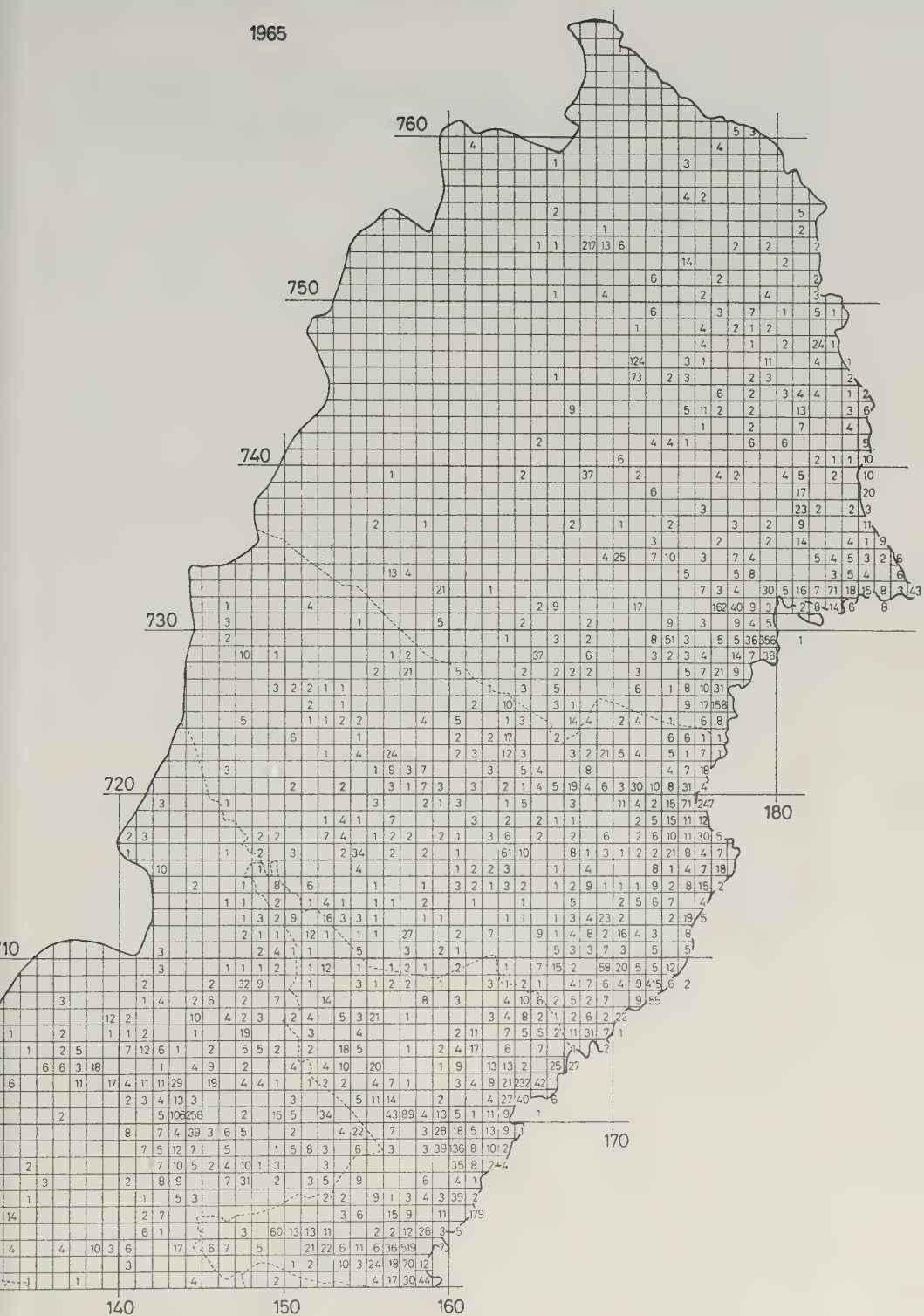


Figure 4.6. Square net map of population distribution in northern Sweden in 1965. The population is given in hundreds per 100 km².

tion equations between the local coordinate system and that used by the land use map. These equations could then be used to transform any coordinate given in the first system to a coordinate in the second. The dots were then located using the coordinatograph and the results used as data for the KOGRAF program. This program determines and uses the transformation equations to calculate the coordinates of the dots in the coordinate system of the land use map. The output from KOGRAF is data of the type required by the square net program NORK. The result of these operations was the square net maps shown in figures 4.3 and 4.4.

4.4 The distribution of population in Sweden in 1965

A number of revisions of Sten de Geer's map have been carried out on different occasions (Pålsson 1958, Hedbom, Norling & Pålsson 1964, Hedbom & Norling 1968). These revisions have reflected the population decrease in country areas in a reduction of the number of dots in the parishes affected. This was done without any knowledge of where in the parish the population reduction has occurred. However better information regarding the population of built-up areas is now available. The later versions of this map show the distribution of population in built-up areas much better than the original.

The latest version of this population dot map shows the distribution of population in 1965 (Hedbom & Norling 1968). It is, of course, completely possible to use this version as the basis for a square net map using the method described in the previous section. As the 1965 version is only a revision of the 1917 version this square net map would not be completely independent of earlier map constructions. It is therefore desirable to use a new map of population distribution that is totally independent of Sten de Geer's map.

A sufficiently accurate map of southern and central Sweden can be constructed using coordinate based parish data in conjunction with the NORK program. The resulting map is shown in figure 4.5. This map gives a relatively good picture of the distribution of population especially in those areas where parishes are small, such as Västergötland, Gotland and southwestern Skåne. In other areas such as Småland there are many squares without any population, which is a result of the fact that the parishes are larger than 100 km². These squares are of course not devoid of people. It is of course possible to distribute the population of country areas amongst these empty squares, in a similar way to that used by Sten de Geer. His method results in a map which appears to be correct but which is actually

as incorrect as that map shown in figure 4.5. The size of square used in analysing the map would in most cases be larger than 100 km², thus reducing considerably the number of empty squares.

The use of parish data for northern Sweden gives a much too inaccurate result as the parishes are extremely large. Sweden is also divided into electoral districts, and in the north of Sweden these electoral districts are much smaller than the parishes. A local authority area can be divided into two or more electoral districts if it has a large population or a large area. This has resulted in many electoral districts in northern Sweden with very small populations. It can be mentioned that the 6 most northerly counties contain about 130 electoral districts with less than 100 inhabitants. The polling station usually lies at a central point in the district when attention is paid to the distribution of population within the district. An accurate population distribution map can be produced using coordinate based population data organized in electoral districts. Other investigations have also shown that electoral district data is highly suited to the needs of regional analysis (Bylund & Weisglass 1970). The square net map shown in figure 4.6 is the result of the processing of population data at the electoral district level obtained from Erson & Weisglass (1969). This processing has been done using the square net program NORK. For ease of presentation the number of persons resident in each square has been rounded off to the nearest 100.

4.5 The accuracy of the different maps

It is of course exceedingly difficult to decide which of the 6 maps shown in figures 4.1 to 4.6 is most true to the situation it purports to represent. In its way the map shown in figure 4.1 is the most accurate as it was constructed using population data for very small units, i.e. villages. Inaccuracies were introduced into the map by the introduction of class divisions and the use of colours to indicate which class is referred to. This fact makes it more likely that figure 4.6 shows its situation in the most accurate manner. The data originates from a fairly small unit, the electoral districts, also the data is only rounded off to the nearest hundred. The other 4 maps are all based on data from the parish level and should therefore be about equally accurate. One can summarize the discussion by saying that all 6 maps are accurate enough to be used in the analysis of inter-regional conditions. At lower levels in the planning hierarchy, for example inside the region, more accurate maps must be used, maps with a grid square of size 1 km² or 1 ha are more suitable. In such cases individual real estates or city blocks must be used as the data base.

5 A mathematical location model

5.1 The multiple location problem

The basic assumption is that a population is spread over a surface and that one is to determine the locations of a number of facilities $F_1, F_2, F_3, \dots, F_N$ which are to serve this population. The locations of $F_1, F_2, F_3, \dots, F_N$ must be found in such a way that the total cost for clients (customers, patients) to visit or use the service facilities is as low as possible. It is assumed that each client has exactly the same need for the service as the other clients, and that he uses the service facility which lies nearest his home. The concept of cost has here a very wide meaning. It can naturally be a pure cost expressed in money, but it can just as well be the time it takes for a client to reach his nearest service facility, or the risk of him being involved in an accident on the way to or from the facility. If this cost is proportional to the transportation distance, then one gets the same result simply using this distance.

A median is defined as the point for which the total transportation distance is a minimum. Thus the general multiple location problem is identical to the problem of determining the position for one or several median points (Bunge 1966). In most cases, however, one has a concave freight cost function which means that there is a decreasing cost per kilometre with increasing distance from the source. As demonstrated in the next two chapters of this book, the concavity of the freight cost function affects the positions of the service facilities insignificantly. Thus, it is quite acceptable to consider the location problem as a determination of multiple medians, and thereby consider only the question of transportation distance. On the other hand, there is no problem in constructing a multiple location model in such a way that even real carriage tables or nonlinear freight cost functions can be used.

5.2 Locating one single activity

The function for finding the total cost incurred by clients when visiting the nearest service facility, in the case of two or more facilities, is mathe-

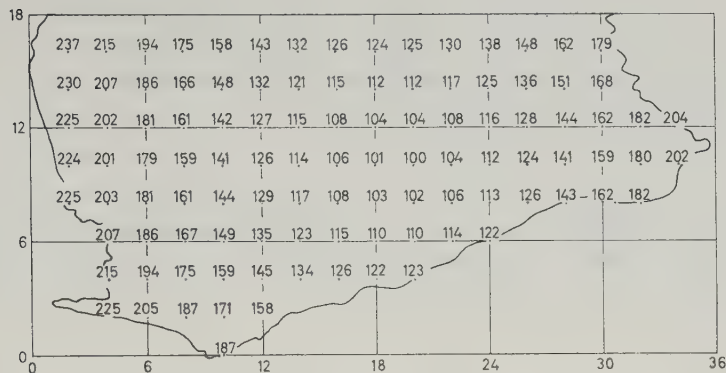


Figure 5.1. A numerical map of southern part of Ghana showing the variation of the total transportation cost function.

atically very complicated, even if it is well defined. However, it is not difficult, for the case with only one facility, to imagine what the function $f(X_{F1}, Y_{F1})$ looks like or to express it in the form of a formula. Let the facility be denoted by $F1$. Thus, in formula 5.1 (X_{F1}, Y_{F1}) are the coordinates

$$f(X_{F1}, Y_{F1}) = \sum_{i=1}^L \sum_{j=1}^W P_{ij} \text{ COST}(D_{ij,F1}) \quad (5.1)$$

for $F1$. The population at point (i,j) is P_{ij} , but since data are usually given in the form of a square net map, P_{ij} is also the number in square (i,j) on the map. The number of rows in the square net map is L , and W is the number of columns. The distance between point (i,j) and $F1$ is denoted $D_{ij,F1}$, and $\text{COST}(D_{ij,F1})$ denotes the cost for a person living in square (i,j) to use facility $F1$.

Function $f(X_{F1}, Y_{F1})$ is easily given in numerical form, as is apparent from the following example. Figure 5.1 is a numerical map of the southern part of Ghana and shows how the total transportation cost function $f(X_{F1}, Y_{F1})$ varies for different positions of the service facility. The function value at each grid point is divided by the minimum value of the function and multiplied by 100. As can be seen from the figure, the function is convex and shaped like a bowl. Thus, it has only one minimum. The function values are calculated in exactly the same way as used in constructing a population potential map. However, the freight cost or time expenditure rises, of course, with increasing transportation distance. In potential calculations interaction (which decreases with increased distance) is used instead of freight cost. But from a purely technical point of view calculating the total transportation distance is identical to the corresponding

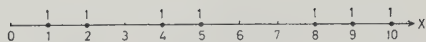


Figure 5.2. An example of a simple 1-dimensional population distribution.

potential calculations (Nordbeck 1964). A map of the type shown in figure 5.1 has a distinct advantage over the corresponding potential map, because one can see directly where the function $f(X_{F1}, Y_{F1})$ has its minimum and thus where the facility should be located. It is also easy to evaluate how uneconomical the other positions are.

5.3 Locating several facilities in a 1-dimensional population distribution

The function for total transportation costs becomes very complicated when one attempts to locate more than one facility within an area. It is nevertheless considerably simplified if it is assumed that the actual population distribution is 1-dimensional; for example, when the population is distributed along a river. Figure 5.2 shows an example of such a simple distribution. If one facility $F1$ is to be located along the line, one receives the function $f(X_{F1})$ given in numerical form in figure 5.3. One can see immediately that the minimum for $f(X_{F1})$ is located at $X_{F1} = 5$. In accordance with Bunge (1966) this point is the median point. The function is convex and easy to treat. Purely mathematically, it is expressed as formula 5.2. In this formula P_j is the number of persons living at point j . The distance

$$f(X_{F1}) = \sum_{j=1}^W P_j \text{ COST}(D_{j,F1}) \quad (5.2)$$

between j and facility $F1$ is $D_{j,F1}$, and $\text{COST}(D_{j,F1})$ is the freight or transportation cost function, i.e. the cost to transport an individual the distance $D_{j,F1}$. As usual X_{F1} is the coordinate for $F1$. W is the number of points on the line corresponding to W , the number of columns in the 2-dimensional square net map.

If two units $F1$ and $F2$ are to be located one gets the function $f(X_{F1}, X_{F2})$ according to formula 5.3. Here X_{F1} and X_{F2} are the coordinates for $F1$

$$f(X_{F1}, X_{F2}) = \sum_{j=1}^W P_j \min(\text{COST}(D_{j,F1}), \text{COST}(D_{j,F2})) \quad (5.3)$$

and $F2$ respectively, and $D_{j,F2}$ is the distance between point j and $F2$. Thus one must choose the smaller of the two $\text{COST}(D_{j,F1})$ and $\text{COST}(D_{j,F2})$

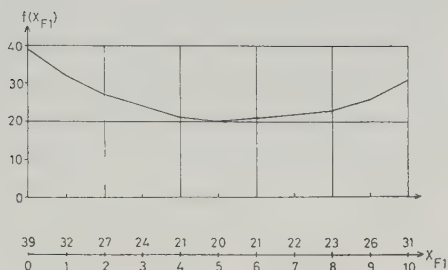


Figure 5.3. The function $f(X_{F1})$ given in numerical form and as a curve in a diagram.

when determining the contribution from point j to total transportation cost $f(X_{F1}, X_{F2})$.

In order to express $f(X_{F1}, X_{F2})$ in numerical form, it is necessary to use a coordinate system with a X_{F1} axis and a X_{F2} axis. If the linear population distribution of figure 5.2 is used, the result is the numerical map of figure 5.4. Here it can be seen that $f(X_{F1}, X_{F2})$ has its minimum at $X_{F1} = 9$ and $2 \leq X_{F2} \leq 4$.

The values from figure 5.3 are found again on the diagonal of the numerical map; that is, the placement of two facilities at one and the same point is the same, from the viewpoint of transportation cost, as placing only one facility at the point in question. The total transportation cost decreases significantly when working with two facilities instead of one, going from 20 to 8 units. The reduction continues for the location of three facilities. In this case, the total transportation cost function $f(X_{F1}, X_{F2}, X_{F3})$ has a minimum value of 4 units. This minimum is achieved when $1 \leq X_{F1} \leq 2$,

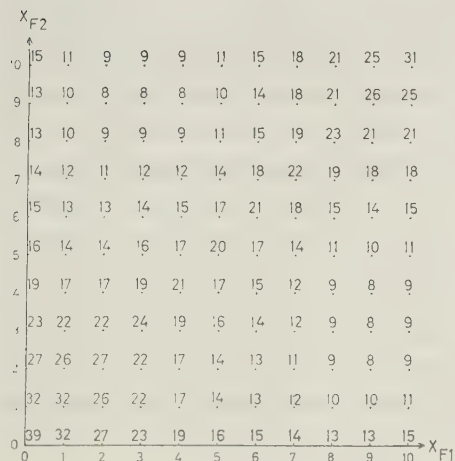


Figure 5.4. The function $f(X_{F1}, X_{F2})$ given in numerical form. Data is the population distribution in figure 5.2.

$4 \leq X_{F2} \leq 5$ and $X_{F3} = 9$. This function is also symmetric, i.e. X_{F1} , X_{F2} and X_{F3} can change places with each other in all combinations.

In order to illustrate the function $f(X_{F1}, X_{F2}, X_{F3})$, in the same way that $f(X_{F1}, X_{F2})$ was illustrated in figure 5.4, it is necessary to have a coordinate system with three axes. In terms of drawing a graphical illustration, this is much too complicated. Nevertheless, it is quite possible to let one of the facilities, say $F1$, have a fixed position and then let the other two vary. Thus, the diagram would be of the same type as figure 5.4. After this, $F1$ can be moved one step and a new "slice" can be obtained. This procedure is repeated for all possible positions for $F1$, thus obtaining a total of 10 diagrams which can be stacked together to form a cube. With the aid of these diagrams, it is then quite easy to determine a minimum value for $f(X_{F1}, X_{F2}, X_{F3})$ and find which values of X_{F1} , X_{F2} and X_{F3} yield this minimum.

5.4 Locating two facilities in a 2-dimensional population distribution

The problem now is to locate optimally two facilities, $F1$ and $F2$, in a given area. In other words, one wants to minimize the function $f(X_{F1}, Y_{F1}, X_{F2}, Y_{F2})$ where (X_{F1}, Y_{F1}) and (X_{F2}, Y_{F2}) are the coordinates for $F1$ and $F2$ respectively. Formula 5.4 is the mathematical form of the function $f(X_{F1}, Y_{F1}, X_{F2}, Y_{F2})$.

$$f(X_{F1}, Y_{F1}, X_{F2}, Y_{F2}) = \sum_{i=1}^L \sum_{j=1}^W P_{ij} \min(\text{COST}(D_{ij,F1}), \text{COST}(D_{ij,F2})) \quad (5.4)$$

It contains similar symbols and notations to formulas 5.1 to 5.3 and is, in principle, only a generalization of them. Unfortunately, it is impossible to represent $f(X_{F1}, Y_{F1}, X_{F2}, Y_{F2})$ pictorially, since this would require a coordinate system with 5 axes. If the numerical map technique is used instead, it is possible to use only 4 axes, but even such a coordinate system is too complicated for displaying the function. It is, however, possible to fix the position for $F1$ by setting $X_{F1} = Y_{F1} = 1$ and then calculate the function value when X_{F2} and Y_{F2} vary according to the method described in section 5.2. A numerical map of the type shown in figure 5.1 is then obtained. Next the position for $F1$ can be changed by setting $X_{F1} = 2$ and repeating the calculations for this case. Each new position for $F1$ results in another numerical map. A total of 162 maps were produced for the example from Ghana since $L = 9$ and $W = 18$. Using these maps, there is no great difficulty in determining the coordinates (X_{F1}, Y_{F1}) and (X_{F2}, Y_{F2})

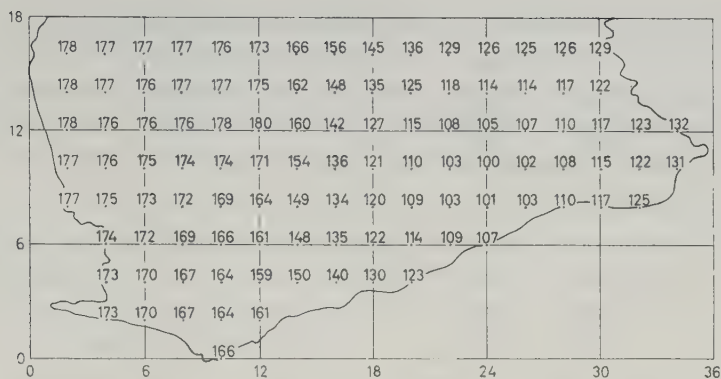


Figure 5.5. Two facilities, F_1 and F_2 , are to be located within southern Ghana. The function $f(X_{F_1}, Y_{F_1}, X_{F_2}, Y_{F_2})$ can be illustrated by means of a great number of numerical maps. The map in this figure is just one of these maps. It shows how the value of the function varies when the location of F_1 is fixed ($X_{F_1} = Y_{F_1} = 12$).

which minimize the function $f(X_{F_1}, Y_{F_1}, X_{F_2}, Y_{F_2})$, i.e. the total transportation costs of the clients.

Figure 5.5 shows one of the 162 numerical maps of southern Ghana described above. It is the one containing the absolute minimum of the function $f(X_{F_1}, Y_{F_1}, X_{F_2}, Y_{F_2})$. The fixed position facility has coordinates (12,12), i.e. $X_{F_1} = Y_{F_1} = 12$. One knows that, since the highest value for $f(X_{F_1}, Y_{F_1}, X_{F_2}, Y_{F_2})$ is obtained when both F_1 and F_2 lie at the same point. This maximum value is 180 for the map shown in figure 5.5. By using this maximum value, it is possible to determine X_{F_1} and Y_{F_1} . The other facility should be placed at point (24,10), i.e. $X_{F_2} = 24$ and $Y_{F_2} = 10$. Thus the function assumes its absolute minimum value at this point. In order to make the map easier to read, the function values have been expressed as percentages of the minimum value.

That the location of facility F_1 at point (12,12) and F_2 at (24,10) really does produce a minimum for $f(X_{F_1}, Y_{F_1}, X_{F_2}, Y_{F_2})$ can easily be checked by going through all 162 maps and seeing if there is a lower value of the function. Since this is not the case it can be concluded that the function's absolute minimum point is (12,12,24,10), or at least in the vicinity of this point. A technique to show that this point is a minimum point is to approach it from different directions and verify that its value is lower than those for all other points in the near vicinity of it. This technique is illustrated most easily by an example from the map of Ghana. In the middle of figure 5.6 lies a square extracted from the map of figure 5.5. It contains 9 smaller squares covering the area for which $22 \leq X_{F_2} \leq 26$

and $8 \leq Y_{F2} \leq 12$. Similarly, 8 other maps have been extracted and placed around the first one so that the corresponding intervals for X_{F1} and Y_{F1} , i.e. $10 \leq X_{F1} \leq 14$ and $10 \leq Y_{F1} \leq 14$, are also included. It should be observed that a side of a square of the grid map is equal to 2 coordinate units. Thus positions are always given in even coordinates.

Using figure 5.6, the point (12,12,24,10) can be approached from several different directions. Of the other points, which lie in the near vicinity of this point, none have a lower functional value. Thus it is the minimum point. This, of course, is true only under the conditions that the function is defined simply at the grid points; in this case only at those points which have even whole numbers as coordinates. On the other hand there is no reason for not using a finer grid net and thereby determine the minimum point with greater precision.

5.5 The general problem of locating multiple facilities

Formula 5.4 can very easily be extended to encompass the general case for locating several facilities in a given area. The result is formula 5.5. The symbols and notations in this formula correspond to that of formula 5.4.

$$f(X_{F1}, \dots, Y_{FN}) = \sum_{i=1}^L \sum_{j=1}^W P_{ij} \min_{k=1..N} (COST(D_{ij,Fk})) \quad (5.5)$$

Thus X_{Fk} and Y_{Fk} are the coordinates for facility Fk , and $COST(D_{ij,Fk})$ is the cost for a person living in square (i,j) to use facility Fk . At this point, the formula should need no further explanation.

5.6 The multiple location algorithm

Once the function $f(X_{F1}, \dots, Y_{FN})$ has been defined the problem of how the point at which it is at a minimum is to be found, arises. Unfortunately the function is, mathematically, far too complicated to allow the use of ordinary methods for this purpose. A numerical method must therefore be used to locate the minimum point and the minimum value. Several such methods are described in the literature, (Gilmore 1962, Kurtzberg 1962, Cooper 1963, Törnqvist 1963, Kuehn & Hamburger 1963, Armour & Buffa 1963, Casetti 1964, Maranzana 1964, Nugent, Vollman & Ruml 1968, Scott 1971, Nordbeck & Rystedt 1971). Amongst these can be mentioned an algorithm that Cooper (1963) and Casetti (1964) developed

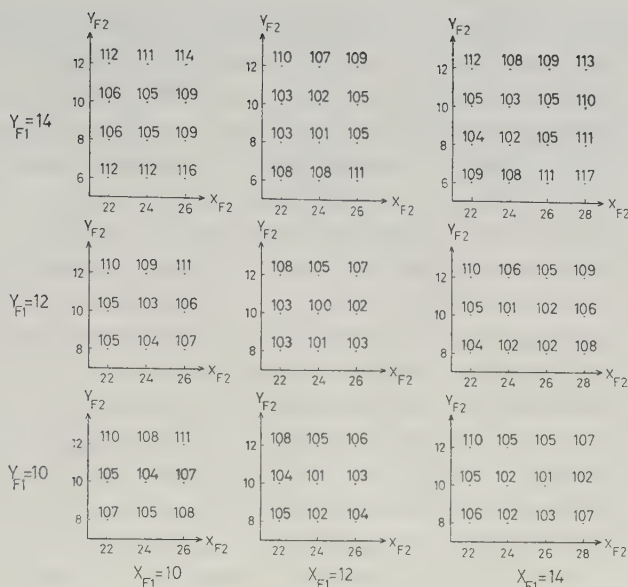


Figure 5.6. Numerical maps showing the value of the function $f(X_{F1}, Y_{F1}, X_{F2}, Y_{F2})$ when the point $(12, 12, 24, 10)$ is approached from different directions.

independently of one another. The approach divides up the population into N groups, thus N umlands are created. The next stage works with every umland independently, an optimum location for a facility being chosen within each umland, using a technique similar to that described in section 5.2. The boundaries of the umlands are then adjusted in such a way that the individual customer is referred to the facility lying nearest his home. This operation results in N new umlands. The whole operation is then repeated, until a stable solution is obtained, with no alterations to either umlands or facility locations occurring between two consecutive iterations.

Törnqvist uses a slightly different method to Cooper. He begins by giving N different locations for the facilities. Each customer is then referred to the facility lying nearest his home, in this way N umlands are obtained. The value the function assumes, with the facilities located at the N initial locations, is then calculated. This is followed by the calculation of the function value at points in the vicinity of the initial location of the first facility. If a point which results in a lower value of the function is discovered the first facility is moved to this point. The procedure is then repeated with each facility. The investigations are halted when a stable solution is obtained, i.e. when any move in the position of a facility does not result in a lower value of the function.

The function $f(X_{F1}, Y_{F1}, \dots, X_{FN}, Y_{FN})$ contains $2N$ variables, viz. $X_{F1}, Y_{F1}, X_{F2}, Y_{F2}, \dots, X_{FN}, Y_{FN}$. Nordbeck and Rystedt (1971) give initial values for these variables, in the same manner as Törnqvist. It should perhaps be pointed out that the initial values can all be the same, but that shorter calculation time is obtained when the initial values lie in the vicinity of the optimum value. Nordbeck and Rystedt begin by locating a minimum point for the function, with all other variables than X_{F1} held fixed. When a minimum value of the function has been found a new value of X_{F1} , to be used in the continued calculations, has also been found. In the next stage of the calculations all the variables other than Y_{F1} , are held fixed and a new minimum value for the function, with a corresponding new value of Y_{F1} , is obtained. This new value of Y_{F1} is used in the continued calculations. A further variable is chosen for variation and a corresponding new value obtained. This procedure is repeated with all the $2N$ variables. The result of this first iteration is a collection of N locations all of which are improvements on the given initial variables' values. The exception, of course, being if the initial locations describe an optimum arrangement. The new set of locations are used as the initial values for iteration number 2. The iteration process is repeated until a stable solution is obtained, with 2 successive iterations leading to the same result. In this way a minimum point for the function $f(X_{F1}, \dots, Y_{FN})$ is obtained. In most cases only 3 or 4 iterations are necessary to locate this point. The algorithm described in simple terms in this section is used in the NORLOC program (Nordbeck & Rystedt 1971). It is, in comparison with other algorithms, relatively fast. Thus NORLOC was used in the location calculations which are presented in the remainder of this report.

6 The location function and local minima

When only one facility was to be located in an area, it was seen that the total transportation cost $f(X_{F1}, Y_{F1})$ was a convex, bowlshaped function with only one minimum. This was true as long as the freight cost function was linear or convex (Cooper 1967). An interesting and important question in this connection is whether the general function $f(X_{F1}, \dots, Y_{FN})$ as defined in formula 5.5, also is convex. Unfortunately, from a mathematical point of view, this is an extremely difficult question to handle. Theoretically it should be possible to differentiate $f(X_{F1}, \dots, Y_{FN})$ twice, and then examine the sign of the second derivative. The function is defined as a sum of a number of terms, and it is not difficult to differentiate such a sum. However, the terms are very complicated. They, in their turn, are functions of distance and each such term is the least of several such functional values. Certainly, at this point, it is impossible to handle $f(X_{F1}, \dots, Y_{FN})$ according to ordinary mathematical rules. Nevertheless, it is possible to tackle the problem purely numerically according to the technique taken up in section 5.4 and see if $f(X_{F1}, \dots, Y_{FN})$ has only one minimum.

6.1 1-dimensional population distribution and local minima

Since the general function $f(X_{F1}, \dots, Y_{FN})$ is so complicated to work with, only the problem of locating two facilities $F1$ and $F2$ will be considered here. Thus the problem is to locate optimally two facilities for a given population distribution. At the same time it must be determined if there is more than one solution to this problem.

As usual, it is easier to work with a 1-dimensional population distribution than with a 2-dimensional one. A simple 1-dimensional distribution is shown at the bottom of figure 6.1 with 4 people living at point (0), 8 at

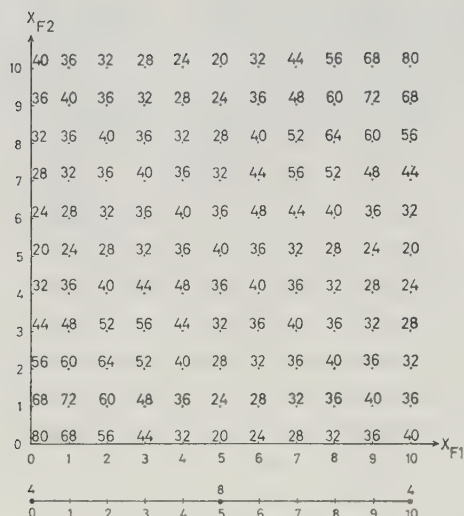


Figure 6.1. The total travel distance $f(X_{F1}, X_{F2})$ has two points of minimum namely (0,5) and (5,10) if the population distribution illustrated in the bottom of the figure is used as data.

point (5), and 4 at point (10). The function $f(X_{F1}, X_{F2})$ is shown above the figure in the form of a numerical map. It is easy to see that its minimum value is equal to 20 and occurs when $X_{F1}=0$ and $X_{F2}=5$. Since the function is symmetrical, this same minimum value is also obtained for $X_{F1}=5$ and $X_{F2}=0$. However, upon a closer examination of figure 6.1, one finds that the function assumes the same minimum value even for $X_{F1}=10$ and $X_{F2}=5$ and for $X_{F1}=5$ and $X_{F2}=10$. Thus there are two different, but equally valid solutions to this location problem. Moreover, if 5 people

Figure 6.2. The population distribution in the bottom of the figure is used as data in the function $f(X_{F1}, X_{F2})$. The function has a local minimum in (0,5) and an absolute one in (5,10).

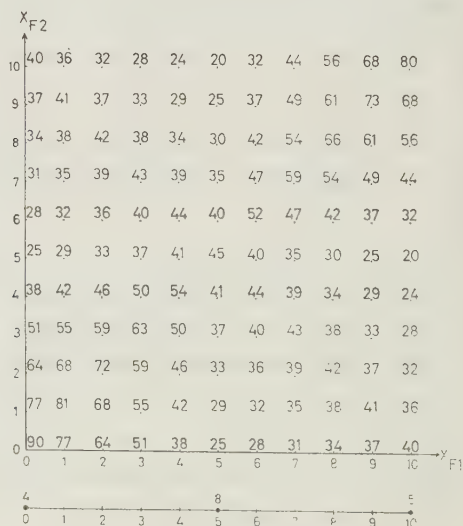


Figure 6.3. A simple population distribution used as input data in the function $f(X_{F1}, Y_{F1}, X_{F2}, Y_{F2})$ in the examples in the figures 6.4—6.8.

8	0	0	0	0	0	29	0	0	0
7	0	0	0	0	111	131	117	0	0
6	0	0	11	27	77	137	145	31	0
5	0	0	19	21	123	97	119	16	0
4	0	0	17	82	23	73	0	13	0
3	0	0	56	41	51	59	71	14	87
2	0	89	109	45	5	2	113	83	67
1	7	103	42	53	1	3	79	101	0
0	0	1	2	3	4	5	6	7	8

lived at point (10), instead of only 4 as in the example, a minimum value (figure 6.2) equal to 25 would have been obtained when $X_{F1}=0$ and $X_{F2}=5$, as opposed to the minimum value of 20 for $X_{F1}=5$ and $X_{F2}=10$. Thus, the example has a local minimum at the point (0,5) and an absolute or global minimum at (5,10).

6.2 Local minima and 2-dimensional population distribution

As was noted in the previous section, and in figures 6.1 and 6.2, the function $f(X_{F1}, X_{F2})$ is not always convex since it can have at least one local minimum. Thus, it is extremely unlikely, and in fact virtually impossible, that the more complicated function $f(X_{F1}, Y_{F1}, X_{F2}, Y_{F2})$ is always convex. The following empirical examples show that this function can have local minima.

6.2.1 Linear freight cost functions

Based upon the population distribution given in the population square net map of figure 6.3, the problem is to locate optimally two facilities $F1$ and $F2$, i.e., minimize the function $f(X_{F1}, Y_{F1}, X_{F2}, Y_{F2})$ defined by formula 5.5. To start with, it can be assumed that the freight cost function is linear, so that the function $COST$ is of the type shown in formula 6.1 with the exponent $B=1$. In this formula D is the distance $D_{ij,F1}$ between square (i,j) and facility $F1$. A is a loading and unloading cost (or if one

$$COST(D) = A + CD^B \quad (6.1)$$

so wishes, a terminal cost), and C is a constant which can be interpreted as a per kilometre cost. It is a reasonable assumption that the terminal cost A is the same for all people, and that it is independent of the transportation distance, so one can omit it from optimization calculations.

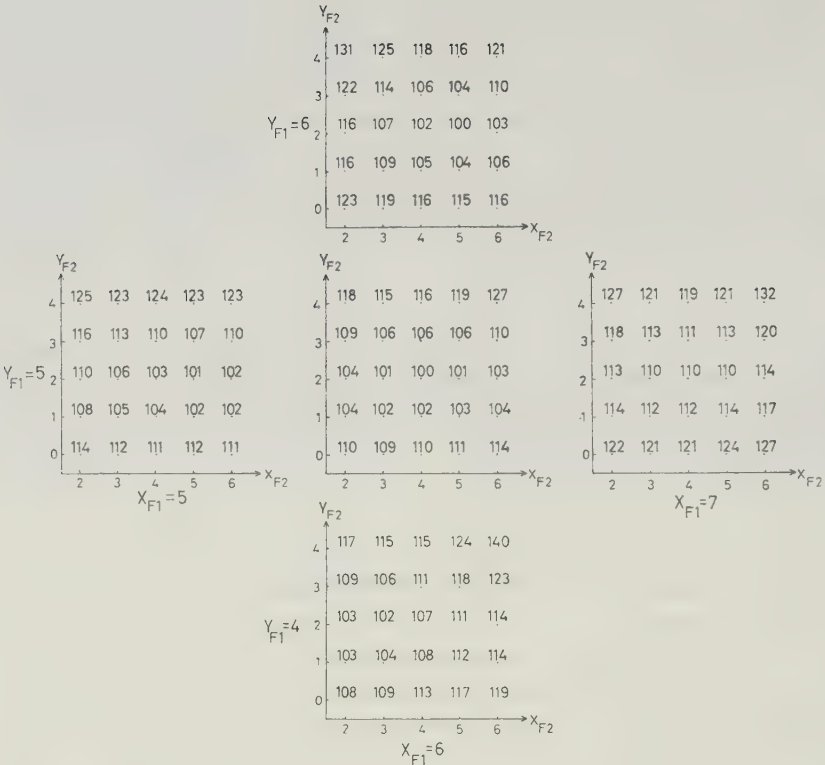


Figure 6.4. Numerical maps showing the value of the function $f(X_{F1}, Y_{F1}, X_{F2}, Y_{F2})$ when the point (6,5,4,2) is approached from different directions.

By the same reasoning, the kilometre cost C in formula 6.1 can also be deleted. Thus, the same positions are obtained for both of the facilities $F1$ and $F2$ whether A and C are included in the calculations or not. It is very easy to correct the final result afterwards by simply multiplying it by the constant C and adding the product PA , where P is the total number of people in the actual area. Under these conditions, it is sufficient to minimize the total transportation distance under the assumption that the freight cost is linear even if it is easy to determine a value for $COST$ according to formula 6.1 or to find it in a table. As this will be done a large number of times, it can result in a significant extension of the calculation time. It is therefore convenient to delete the constants A and C from the actual computing, and then adjust the results afterwards.

An optimal solution for both facilities $F1$ and $F2$ with respect to the population distribution of figure 6.3 results in $F1$ being located at point (6,5) and $F2$ at point (4,2). Figure 6.4 presents a series of numerical maps

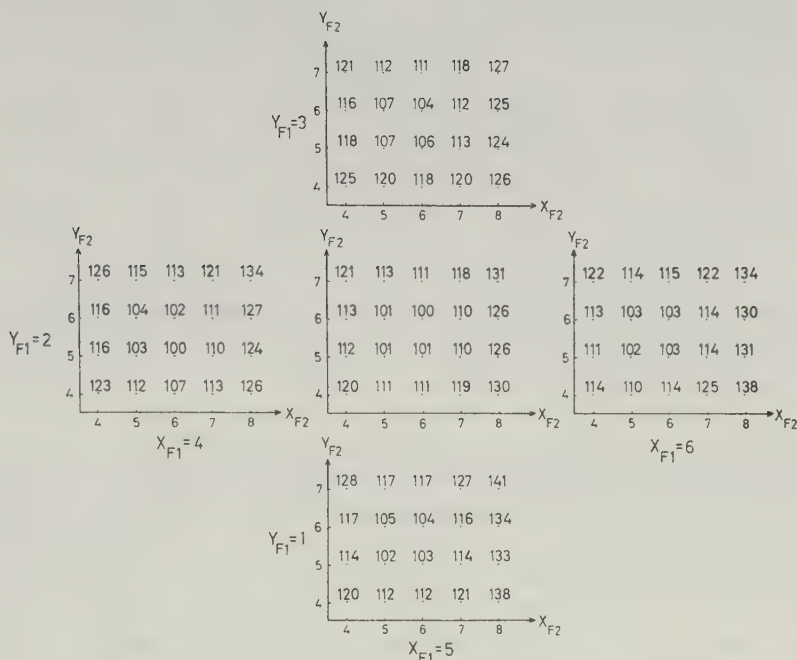


Figure 6.5. Numerical maps showing the value of the function $f(X_{F1}, Y_{F1}, X_{F2}, Y_{F2})$ when the point $(5, 2, 6, 6)$ is approached from different directions.

giving values for $f(X_{F1}, Y_{F1}, X_{F2}, Y_{F2})$ as percentages of its minimum value as it nears these two points from different directions. As can be seen, the function assumes a minimum value for $X_{F1}=6$, $Y_{F1}=5$, $X_{F2}=4$ and $Y_{F2}=2$, or, in other words, at the point $(6, 5, 4, 2)$.

Figure 6.5 shows the values when $F1$ and $F2$ approach the points $(6, 6)$ and $(5, 2)$ respectively. Again, the function assumes a minimum value for $X_{F1}=6$, $Y_{F1}=6$, $X_{F2}=5$ and $Y_{F2}=2$. However, this point $(6, 6, 5, 2)$ lies rather close to the point $(6, 5, 4, 2)$ and it is possible to suspect them as belonging to the same minimum, even if the multiple location algorithm considers them as two different minima when whole numbers are used in the calculations. By combining the numerical maps for these two points (figures 6.4 and 6.5), one can see that a significantly lower value is not obtained when these two points are approached. Thus, the use of a finer meshed net in the vicinity of these two points is probably justified. The functional values are not only to be calculated at whole number points, but also at the half, quarter and finer fractional points. If this process is extended down to 16ths, only one minimum is obtained in the area, namely at the point $(5.50, 5.50, 5.00, 1.94)$. If these coordinates are rounded off

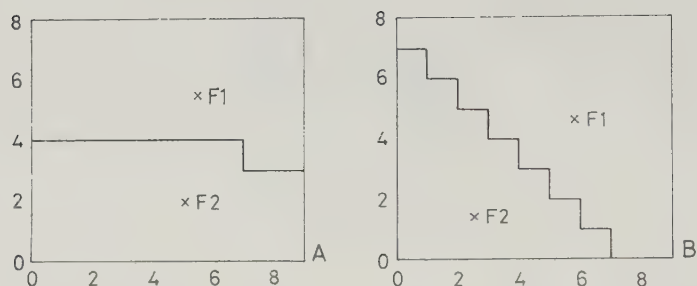


Figure 6.6. The locations and the umlands of the two facilities $F1$ and $F2$ when the grid map in figure 6.3 is used as input data in the location program. The exponent B of the carriage function is equal to 1.0.

to whole numbers, the point $(6,6,5,2)$ is obtained again and can thus be considered as the real minimum point, while $(6,5,4,2)$ is a fictive minimum point. Thus, the suspicion that points $(6,5,4,2)$ and $(6,6,5,2)$ belonged to the same minimum, i.e. that one of them was a fictive minimum, is shown to be correct. The positions $F1 = (5.50, 5.50)$ and $F2 = (5.00, 1.94)$ have been placed in figure 6.6 A together with the corresponding umlands. These positions $F1$ and $F2$ are, however not the only ones which minimize $f(X_{F1}, Y_{F1}, X_{F2}, Y_{F2})$ when a linear freight cost function is used. The point $(5.87, 4.50, 2.50, 1.50)$ is also a minimum point. Furthermore, this point lies far enough away from the other minimum point to make it impossible for them to be the same. Figure 6.6 B shows the positions $F1 = (5.87, 4.50)$ and $F2 = (2.50, 1.50)$, as well as the corresponding umlands. However, the total transportation costs in this case are about 2.5 percent higher than for the case shown in the figure 6.6 A. Thus figure 6.6 B contains a local minimum, while figure 6.6 A has the global, or absolute one.

6.2.2 Concave freight cost functions

If B , in the function $COST = A + CD^B$, is set equal to 0.9, a concave freight cost function is obtained. The total transportation costs of the example are reduced by about 7 percent. Nevertheless, two different solutions to the multiple location problem are still obtained, namely points $(5.50, 5.50, 5.06, 1.94)$ and $(5.81, 4.62, 2.50, 1.50)$. Both of these solutions are shown in figures 6.7 together with the corresponding umlands. It is apparent that figure 6.7 A is nearly identical to figure 6.6 A. The position for $F1$ is exactly the same for both maps, but facility $F2$ is located at point $(5.00, 1.94)$ according to figure 6.6 A and to $(5.06, 1.94)$ according to figure 6.7 A. The distance between these two positions is insignificant. What was noted about figures 6.6 A and 6.7 A holds for figures 6.6 B and 6.7 B

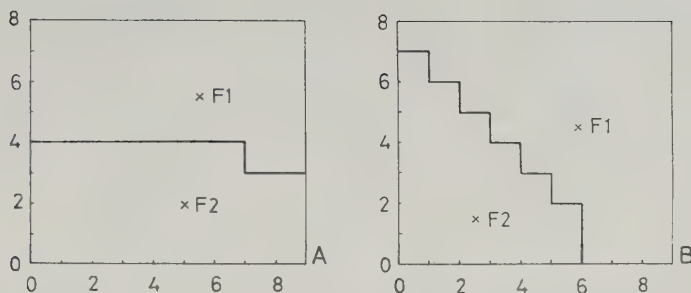


Figure 6.7. The locations and umlands of $F1$ and $F2$ when the exponent B is equal to 0.9.

also. These 4 figures emphasize the point that, by and large, the same locations for $F1$ and $F2$ are obtained whether a linear or concave freight cost function is used.

The total transportation costs for the case shown in figure 6.7 B are about 2 percent higher than those for the minimum in figure 6.7 A. Thus, a local minimum (figure 6.7 B) and an absolute minimum (figure 6.7 A) are also obtained when a concave freight cost function is used. If the multiple location algorithm works with only integers, the fictive minimum (6,5,4,2) is also obtained. This is the same fictive minimum that was obtained by using a linear freight cost function.

6.2.3 Convex freight cost functions

It is even possible to obtain different local minima in the case of a convex freight cost function. If B in the function $COST$ is set equal to 1.3 in the above example, and if only whole number calculations are used three different solutions are obtained for the location problem, namely points (6,5,4,2), (6,6,5,2), and (3,2,6,4). The first two of these points, however, belong to the same minimum. Hence point (6,5,4,2), even in this case, is a fictive minimum point. If 16ths are used in the calculations, the two minimum points (5.50, 5.50, 4.94, 1.88) and (6.00, 4.31, 2.56, 1.62) are obtained. These points lie very close to the corresponding minima points obtained from using the linear and concave freight cost functions $COST$ respectively (see figure 6.8). Thus, there is further support for the notion that the value of the exponent B in $COST$ can vary within reasonable limits without affecting the locations of the facilities $F1, F2, \dots, FN$ and their corresponding umlands.

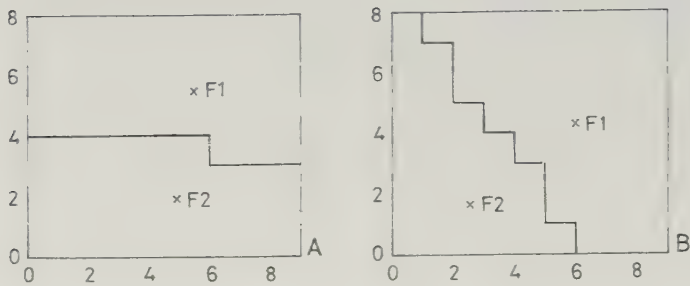


Figure 6.8. The locations and umlands of F_1 and F_2 when the exponent B is equal to 1.3.

Table 6.1 presents the different minimum values and corresponding minima points for the exponents 0.9, 1.0 and 1.3. A relative number for every minimum point is given in the right hand column. This number is simply the quotient of the local minimum value and the corresponding global minimum value. This quotient is then multiplied by 100.

6.3 Local minima and groups of the population distribution

Upon a closer examination of the map in figure 6.3, it can be seen that the population consists of three different groups which can be called the northern, the southwestern, and the southeastern group. The last two groups are about the same size in terms of population and area, whereas the group in the north is larger. As the problem is to locate only two facilities in the square net map area, one group will not have a very central facility. Giving one to the large northern group, it is obvious that the other facility can then be located in either the southeastern or in the southwestern group. Thus two different solutions to the locational problem should be obtained, and the function $f(X_{F1}, Y_{F1}, X_{F2}, Y_{F2})$ should have two different minimum points. If the number of inhabitants in the northern group had been reduced so it was approximately of the same size as the other two groups, three different minimum points would have been obtained.

Summarizing, it has been established that the total transportation cost function $f(X_{F1}, Y_{F1}, \dots, X_{FN}, Y_{FN})$ is not always a convex function with only one minimum. The risk of having one or several local minima is largest when one is dealing with a population which can be divided up into groups well separated from each other in a geographic sense, and when

Exponent	Positions		Total costs	Relative number
	F1	F2		
0.9	5.50,5.50	5.06,1.94	4668	100.0
1.0	5.50,5.50	5.00,1.94	5041	100.0
1.3	5.50,5.50	4.94,1.88	6416	100.0
0.9	5.81,4.62	2.50,1.50	4762	102.0
1.0	5.87,4.50	2.50,1.50	5169	102.5
1.3	6.00,4.31	2.56,1.62	6669	103.9

Table 6.1. Minimum positions and corresponding total costs for different exponents of the carriage function $COST = A + CDB$ used in the locational calculations. The square net map of figure 6.3 was used as input data. The relative number in the column to the far right is 100 times the local minimum value divided by the corresponding absolute minimum value.

the number of such group is larger than the number of facilities to be located.

6.4 Practical planning problems and local minima

It can be seen from the table 6.1 that the function $f(X_{F1}, Y_{F1}, X_{F2}, Y_{F2})$ can have several local minima whether the carriage function $COST$ is concave, linear, or convex. Heuristic locational algorithms, by definition, are devised in such a way that one cannot be certain if the absolute minimum value of the function $f(X_{F1}, Y_{F1}, X_{F2}, Y_{F2})$ has been obtained. It is not even known for certain if a solution is a pure local minimum since it can very well be a fictive one. The problem is compounded when locating more than two units, but in these cases the function is so complicated as to make it very hard to work with, even purely numerically. Even so, the multiple location algorithm can be of enormous help in practical planning problems. As Shen Lin (1965) has noted in another combinatorial and spatial problem, the difference between the absolute minimum value and that value obtained by an algorithm is usually quite small. On the other hand, one must be extremely careful not to exclude the possibility of true global solution by referring to the searchlike calculations of the algorithm.

An example, showing how the mathematical location model embodied in NORLOC can be used in a practical context, is provided by the problem of dividing Sweden into college districts. In this case the problem is further complicated by the presence of local minima. The population of a college region should not be less than 33 000 persons and this requirement has been used as the basis for a division of Sweden into 69 regions. One of these regions is made up of the island of Gotland. It is quite obvious that Gotland with its population of about 54 000 people

	Mean distance
Start	19.1
Minimum 1	17.5
Minimum 2	17.3

Table 6.2. The mean travel distance in kilometres, for all Swedes, from their home to the nearest of the 68 towns having a college (gymnasium). The calculation has been carried out for 3 alternatives, manually chosen locations (start) and with choice of locations performed by NORLOC (minima 1 and 2).

(1965) will even in the future function as an independent region and it has therefore been excluded from this study. Only the 68 regions on the Swedish mainland are discussed here.

One begins by giving the centres of 68 regions as the initial locations required by the NORLOC program. The first stage in the locational calculations, carried out by this program, is the calculation of the value associated with the function $f(X_{F1}, Y_{F1}, \dots, X_{F68}, Y_{F68})$ using the initial locations of the 68 school centres. This value is the total transport distance incurred when every person travels from his home to the college centre nearest his home. This result can be divided by the total population to give the mean distance. Table 6.2 shows the mean distances for the different stages, the initial value being 19.1 km. The NORLOC program then finds a minimum point for the function $f(X_{F1}, Y_{F1}, \dots, X_{F68}, Y_{F68})$, thereby dividing Sweden into 68 regions with a centre for each. In the final arrangement the mean travel distance is reduced to 17.5 km.

The locations chosen for regional centres number 45, 68 and 65 together with their corresponding regions are shown in figure 6.9. The population base in region 68 is 32 900 and this makes it the smallest region. The map in figure 6.9 shows that the centre of region 68 lies very near the centre of region 45. The combination of these two regions would not therefore bring about major problems. Region 65 has a very large area and a division of it into two parts could be considered. The division of a region or the combination of several regions into one means that the regional boundaries require adjustment, which should result in different locations for the regional centres. This is best done by application of the NORLOC program once more. The centre of region 68 was given an initial location at (170, 755), the initial values used for the other 67 were those obtained from the previous application of NORLOC.

The resulting regional arrangement is shown in figure 6.10. A comparison with the results from the first application of the program shows that

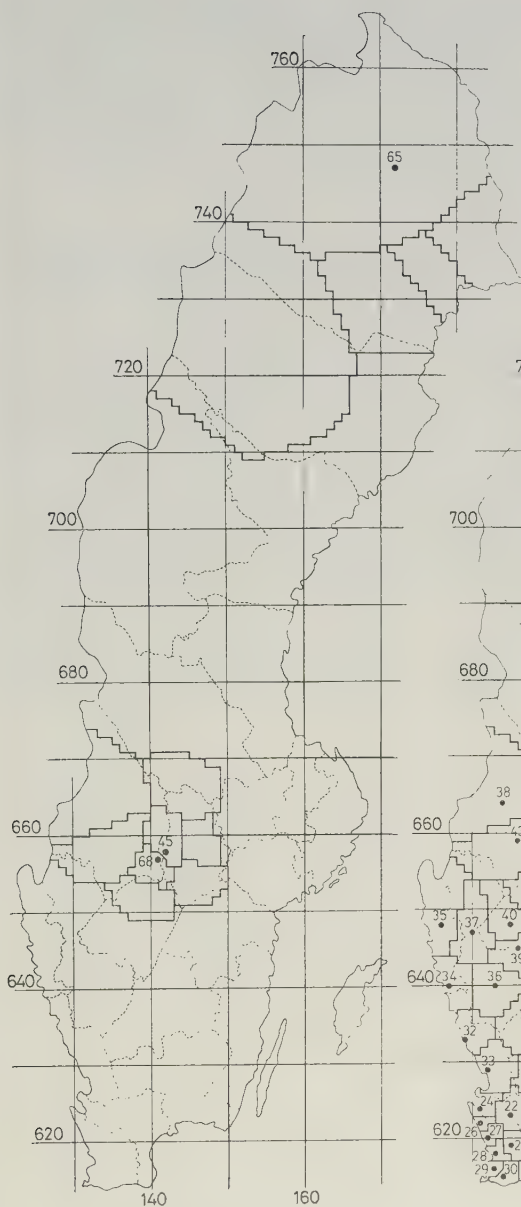


Fig. 6.9

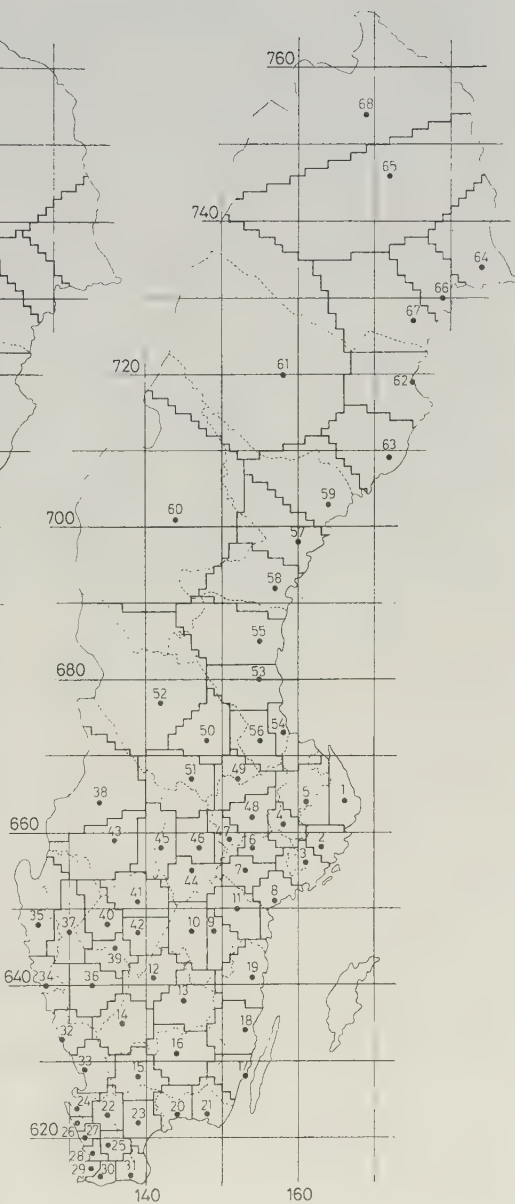


Fig. 6.10

Figure 6.9. The location and umland of facilities 45, 68 and 65, as chosen by NORLOC. The number of facilities located was 68.

Figure 6.10. The location and umland of 68 facilities, as chosen by NORLOC. This arrangement is somewhat better than that shown in figure 6.9, even allowing for the fact that the majority of regions are identical in both arrangements.

region 41 has been extended to the north, region 43 extends further to the east, and the largest part of the previous region 68 has been transferred to region 45. In northern Sweden the previous region 65 is divided into two and its boundary moved somewhat to the south. This alteration causes minor adjustments to regions 61, 67, 66 and 64. The new region 68 is now the smallest one with a population of 30 500.

Table 6.2 shows that the mean distance is further reduced by these changes, it is now 17.3 km. Application of the location model embodied in NOR-LOC has reduced the total distance travelled by nearly 10 %. This reduction does not mean that the minimum value has been attained. It can be seen from figure 6.10 that the centre of region 29 lies quite near the centre of region 30, it would therefore be possible to combine these two regions at the same time as another region is divided. This operation could result in a further reduction in the mean distance travelled. Against this must be placed the risk of regions with too small a population occurring. This is likely to occur when dividing regions with large areas due to the fact that many of the largest regions have populations only a little larger than the minimum. The planner has therefore a difficult decision to make. He can begin with alternative arrangements and see which of them gives an optimum result. This approach reduces the risk that a satisfactory solution is missed.

7 Sensitivity analysis of the location model

Data and parameters must be given to the computer before the location algorithm can be used to solve a given problem. How accurately these must be specified is a question to which an answer is necessary. The model's sensitivity to inaccuracies in the data or parameters must therefore be determined in some way. One way of doing this uses data and parameters which are assumed to be correct. Then, from this data a set of incorrect data is generated and the resulting location arrangements compared.

Amongst the data necessary when applying the location model, is a description of the customers' spatial distribution. This data is given to the NORLOC program in the form of square net maps. Another approach is to describe the population in the form of subpopulations located at points in a coordinate system. Two factors influence the accuracy of a square net map of population distribution; the amount of attention paid to the location of the individuals when the map was constructed and the size of grid used. These two factors influence one another in that the smaller the grid size used the more difficult it is to place the correct number of individuals in each grid square.

The use of parish data in the construction of a square net map, as was done in the case of figures 4.2 – 4.5, prohibits the construction of a fully accurate square net map, other than in the case where large grid squares are used. Maps produced in this way are therefore always inaccurate, and a comparison of the number of persons living in a grid square with the population the map gives will always show a difference. On the other hand if the comparison is carried out over a larger area, covering a number of parishes, then the inaccuracies cancel one another out, i.e. if the population given for one grid square is too large then the population of neighbouring squares is too small. The mean inaccuracy is thus 0. In the case of the maps shown in figures 4.1 to 4.6 the mean inaccuracy, at for example the county level, is equal to 0.

It is considerably more difficult to estimate the distribution of inaccuracies in a square net map than their respective mean value. The following approach can be used to give a rough approximation of the distribution. Nearly 80 % of the Swedish population lived in built-up areas in 1965. The majority of these built-up areas are also the centre of a parish. This assumption is true in the case of the larger built-up areas and in 1965 about 70 % of Swedes lived in built-up areas with a population over 2000. Thus the majority of these 70 % can be assumed to be located in the proper grid square. Residents in other built-up areas and residents in country districts living at the same grid square as that in which the church is located are in the majority of cases allocated to the correct grid square. In northern Sweden data based on electoral districts was used, which is quite accurate as the electoral districts are small. The distribution of inaccuracies in the maps 4.5 and 4.6 is probably somewhere between 5 and 20 %. Also, the larger the grid square used, the less the relative inaccuracy and its distribution.

The stability of the solution relative to inaccuracies in the map of population distribution is the first question investigated as a part of the sensitivity analysis of the location model. The second question is how differences in the grid size chosen for the square net map affect the stability of the solutions produced. A solution is considered to be stable if both correct data and data containing a certain known degree of inaccuracy gives the same or nearly the same locations.

A further example of the data required for the location model is the cost incurred by every customer in travelling from his residence to the nearest facility. These transportation costs can be given in a table or as a function. The degree of accuracy to which these costs are required to be given must also be determined in advance, due to the difficulties involved in obtaining exact costs. Studies showing the degree of inaccuracy in the transportation costs which results in unstable solutions were therefore necessary.

The distance between the customer's home and the facility is measured as the crow flies, for practical reasons. This distance can be calculated from the coordinates of the two points in a very simple manner using Pythagoras' theorem. The actual journey takes place via a network of some kind. The 4th question studied as a part of the sensitivity analysis is therefore the degree to which the solutions are affected by the use of straight line distances instead of distances via a network.

Sensitivity analysis of the location model is a form of numerical proof, in the same way as the treatment of local minimum points is a numerical proof. In the case of the local minimum points the assumption was made

that the function $f(X_{F1}, \dots, Y_{FN})$ was convex, an assumption which allows the existence of only one minimum point. The function was shown not to be convex, by the discovery of a numerical example having two different minimum points. Hence, it was fairly simple to disprove the hypothesis that $f(X_{F1}, \dots, Y_{FN})$ is a convex function, for all values of N and all possible spatial distributions of the customers. If, on the other hand, the hypothesis had been correct it would have been impossible to prove its validity using the same type of numerical proof. However, every time it was shown that an experimental example had only one minimum the probability of the hypothesis being true would have been increased, but still no final proof would have been given. In the same way, the sensitivity analysis of the location model shows that the solutions produced are stable in relation to inaccuracies in the input data, inaccuracies in the transportation costs and use of straight line distances. Thus, one can use a numerical example to disprove a hypothesis but not to prove absolutely its validity. Repeated application of the sensitivity analysis makes the hypothesis more probable but does not provide a final proof.

7.1 Sensitivity to error in the data

The mapping of, for example, population distribution within an area in an absolutely accurate manner is probably impossible, except in the case of a very small area. In order to carry out an accurate mapping it is necessary to know exactly which persons are resident within the district in question, together with the exact location of every person's residence. A further factor is that the distribution of population within an area is changed very quickly, by births, deaths and migration. This means that a population map which was absolutely accurate when it was constructed soon becomes inaccurate. Even in those cases when coordinate based population data of high quality has been used in the construction of a map, inaccuracies creep in. More serious inaccuracies occur when an approximation of the population distribution is used. The location of the whole population of a parish in the village where the church is located, or the spreading out of the population of the parish using a rule of thumb are examples of such approximations. Such a rule of thumb is that the population is evenly distributed over the surface, or that it has the same distribution as that of dwelling houses. When taken over a larger area errors occurring in this way cancel one another out, i.e. their mean value is 0. One can also assume that these errors are normally distributed, an assumption that in many cases is incorrect. When a population is located on the map, at the village in the parish where the church is located, the errors certainly

are not normally distributed. The assumption becomes more reasonable when the distribution of population is assumed to follow that of dwelling houses. Those errors which are the result of changes in the population structure can be assumed to be approximately normally distributed. The assumption that errors are normally distributed over the surface is open to criticism, but is accepted here, for the purpose of carrying out the sensitivity analysis of the model.

7.1.1 An example from Tanzania

A previously published method was used in the analysis of the location model and its sensitivity to error (Gould, Nordbeck & Rystedt 1971). The starting point was a dot map of population distribution in Tanzania (Porter 1968). The information contained in this map was transferred to a square net map. The assumption was then made that the resulting square net map was free of errors. The NORLOC program was used to determine the optimum location of 5 facilities within Tanzania. The next stage involved the generation of a series of normally distributed random numbers having a mean value of 0 and a standard deviation of 0.05. These random numbers were then stored in a matrix having the same dimensions as the square net map. The elements in the matrix were then multiplied with their corresponding elements in the square net map to give a new matrix, the error matrix, containing errors of the desired character. The values contained in the error matrix were then subtracted from the values contained in the square net map to give a new map containing the desired errors. This new map was then used as input data to the location model. The resulting locations of the 5 facilities were exactly the same as those obtained using the correct map. Hence, the location model was insensitive to errors in the input data having a mean value of 0, which were normally distributed with a standard deviation of 0.05.

Then sensitivity analysis was continued using larger standard deviations when generating the random numbers. The standard deviation was increased in steps of 0.05 giving the series 0.05, 0.10, 0.15, 0.20 etc. Only when the standard deviation reached 0.35 was any instability encountered, since two facilities moved 1 position to the west. This new location pattern differed only to a small degree from the original and was also found to be stable in its own right when the standard deviation was increased still further.

The analysis was discontinued when the standard deviation had acquired the value 0.50. This large standard deviation gave errors of such a size that a number of grid squares had a negative population. The probability

that one generates a random number, the absolute value of which is larger than 1.0 is 0.03, when a standard deviation of 0.50 is used. The Tanzania example referred to here contained 200 populated grid squares which should result, in an average case, in 3 grid squares having a negative population. An increase in the standard deviation above 0.50 increases the probability that grid squares with negative populations will occur. The introduction of such errors into the data was felt to be unrealistic.

One can sum up these investigations by saying that the location model is insensitive to reasonable errors in the square net map used as input data, always assuming that the errors are normally distributed. The model gives the same or nearly the same solution to the location problem without any noticeable influence being exerted by errors in the input data. In order to give this statement further weight a sensitivity analysis using data from southern Sweden is also described.

7.1.2 Southern Sweden

Use of the location algorithm in the location of 6 facilities in Sweden, using a square net map with an equidistance of 10 km, divides the country into 6 regions. In the south of Sweden the major region includes the whole of Skåne together with the whole of the counties of Blekinge and Kronoberg, the southern parts of Halland and Kalmar counties and the island of Öland. (The names and the locations of the Swedish counties are given in figure 7.16 on page 246.) Square net maps of population distribution are available for this area, for the year 1965. The grid size used in these maps is 1 km throughout. The region is thus highly suitable for the carrying out of a sensitivity analysis.

The southern Swedish region covers about 50 000 km². A retention of the original map scale would present major difficulties, prohibiting the application of NORLOC, extremely long calculation times would be necessary. Also the majority of computers are not equipped with fast memories sufficiently large to allow the storage of maps of this size. It was therefore found necessary to combine 25 of the original grid squares in order to produce new squares with sides 5 km long. This results in a square net map containing approximately 2000 squares. This map was then used as input data to the location program.

Before a problem can be handled by the NORLOC program it is necessary to give the number of facilities to be located. It has been found that the location of 7 facilities in southern Sweden produces a version of the function, $f(X_{F1}, \dots, Y_{F7})$, which has only one minimum. This number of facilities was therefore chosen for the sensitivity analysis examples.

The method used was exactly the same as that used for the Tanzania example mentioned earlier, with the exception that the values in the incorrect map were calculated directly without the use of an intermediate matrix. The location model was applied 11 times, using the above input data. The first of these used the original map, which was assumed not to contain any errors. The locations for the 7 facilities obtained in this way are shown as solid circles in figure 7.1. When the model was applied for the second time the data was amended to include normally distributed errors. The mean value of these errors was 0 and the standard deviation 0.05. The resulting locations are shown on the map by open circles. In the following applications of the program the standard deviation was increased by 0.05 every time, up to a maximum value of 0.50. An application of the NORLOC program 11 times using data for southern Sweden should have resulted in 11 different locations for each of the 7 facilities. Due to the fact that the solutions are relatively stable with respect to errors in the data the number of locations obtained is considerably less. The first facility *F1* moved only 1 unit, i.e. 5 km, in an east—west direction and took up only 3 positions on the way. Facilities *F2* and *F4* were nearly as stable as *F1*. They moved 5 and 7.5 km respectively in an east—west direction and 2.5 km in a north—south direction. On the way they assumed 4 different positions, as did facility number *F3*, although in this case the total distance moved was 20 km in a southeasterly direction. The other facilities assumed 5 different positions. In all cases the facilities moved around within a circle having a radius of 10 km. On the basis of these results it can be said that errors in the input data affect the locations assumed by the facilities only to a very small degree. The same is true of the umlands. It can be seen from figure 7.1 that the boundaries of the different umlands only move between 5 and 10 km. These movements create a diffused boundary zone between the different umlands. Even so, it is possible to say that the umlands are stable and insensitive to reasonable errors in the data, always assuming that these errors are normally distributed with a mean value of 0.

In the experiment described above the errors were generated using varying series of random numbers. The experiment could of course be repeated using differing series of random numbers. It is probably not at all difficult to generate a series which gives an even better result than that shown in figure 7.1. The greater the number of experiments that are carried out the greater the authority behind the claim that the location model is insensitive to reasonable errors in the data. On the other hand, it is not advisable to carry out a large number of executions of the program and then to calculate the mean position of each facility. If this is done the

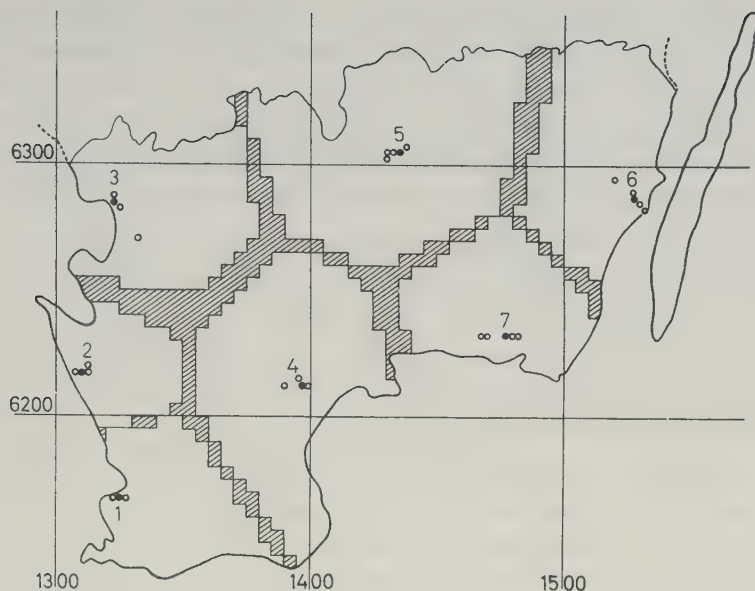


Figure 7.1. The solid circle indicate the optimum locations for 7 facilities in southern Sweden. The input data used was a square net map of population distribution in 1965, with an equidistance of 5 km. The remaining circles indicate the locations generated by the program NORLOC when errors were introduced into the data. The errors were normally distributed with standard deviations 0.05, 0.10, 0.15, ..., 0.50 and with mean values of 0.

random effects are eliminated, the result should therefore be the same as that obtained when using correct data. It should not create any surprise when stable arrangements result from the use of such a procedure. Every execution of the program should be independently analysed. Further executions of the program using different data could be carried out in order to make the claim that the location model is insensitive to moderate errors in the data even more reasonable.

7.1.3. Real estate data — parish data

Correct population maps, based on coordinate data at the real estate or building level, are extremely rare, as has been pointed out earlier. Instead data based on larger areas, for example registration districts and parishes, must be used in conjunction with an approximate distribution of the population. Every grid square in such an approximate map contains a population, but in most cases the size of this population is inaccurate. Over larger areas the errors cancel one another out. The sensitivity of the location model to errors of this type was also investigated.

The population map used when checking the sensitivity of the model to normally distributed errors in the data shows southern Sweden in 1965. This map is of very high quality and was constructed using coordinate based property data at the level of the individual dwelling. Thus, this map can be considered to be completely free of errors. Southern Sweden is therefore highly suitable for this extension of the sensitivity analysis.

Figure 4.5 shows a square net map of population distribution in southern and central Sweden dated 1965. The map is based on coordinate data at the level of the parish. The whole population of each parish has been shown as having the same coordinates as its church. This data can easily be used to construct a map of population distribution in southern Sweden which has a equidistance of 5 km.

Table 7.1 shows the results obtained when using NORLOC and two sets of input data, one set based on a map obtained using data at the level of the parish and one set based on the map obtained when using data at the level of the individual house. The total population in the first case is 1 461 000 and in the second case 1 454 000. The difference between these two total populations can be partly explained by the presence in the first total of persons not registered at any real estate or without any known address. These persons cannot be given coordinates.

As can be seen from table 7.1 nearly exactly the same result is obtained using the approximate map, constructed using parish data, as is obtained using the correct map, constructed using real estate data. The mean distance is practically the same in both cases. Bearing in mind that the whole population of southern Sweden is involved in the calculations the difference between the two values is surprisingly small. In the first case the mean distance is 19.98 km and in the second case it is 19.96 km.

The results from both executions of the program NORLOC show nearly identical locations for the facilities. The majority of the facilities move 0.5 units, (2.5 km) in one direction only. *F7* is the facility which is most affected when parish data is used. It moves 7.5 km in a north—south direction and 2.5 km in an east—west direction. This maximum movement indicates the unimportant nature of the differences between the two sets of data. When the number of populated squares is sufficiently large, in proportion to the number of facilities which are to be located, then the location model is probably insensitive to errors in the input data. It ought also to be observed that the mean error for every parish is 0, due to the fact that the whole population of a parish is given the coordinates of its church.

Facility	Parish data			Real estate data		
	Position	Amount in 1000's	Mean distance in km	Position	Amount in 1000's	Mean distance in km
F1	1327,6167	481	14.1	1325,6167	471	13.6
F2	1312,6217	237	16.3	1310,6217	239	15.9
F3	1317,6287	111	17.3	1322,6285	114	18.8
F4	1397,6212	230	30.0	1397,6212	242	30.1
F5	1437,6307	134	26.1	1435,6305	133	26.6
F6	1530,6285	130	25.5	1527,6287	125	25.0
F7	1470,6235	137	22.7	1477,6232	128	21.2
Total	— —	1461	20.0	— —	1454	20.0

Table 7.1. The optimum locations for 7 facilities in southern Sweden in 1965. Two square net maps have been used as input data to NORLOC, both showing the same area. One map was generated using coordinate based parish data and the other using coordinate based real estate data.

A further question can be raised; do these results from the sensitivity analysis apply at other planning levels? If this is the case then location problems at the local level can be tackled, without the necessity for co-ordinate based real estate data. Hence an urban planner might use co-ordinate based data for blocks instead of data for dwelling houses.

7.2 Sensitivity to changes in the recording grid

The smaller the squares in a square net map, used as input to the location model, the more accurate the results obtained. From this point of view it would be best to organize the input data in a map in grid squares that are as small as possible. Against this must be placed the fact that time required to execute the location program NORLOC is proportional to the number of squares in the map. The choice of grid squares of small size always leads to long execution times, whereas the choice of a larger grid always results in locations being given less accurately. Fortunately an investigation has shown that stable solutions are obtained, even when the size of the grid square is varied within reasonable limits (Gould, Nordbeck & Rystedt 1971). The starting point for this investigation was a population distribution map with grid squares of small size. This map was used as input to the location program and a result obtained, which was assumed to be correct. In the next stage of the investigation new maps of population distribution were constructed by aggregating a number of squares in the original, 4, 16, 64 and so on. Every new map was constructed by doubling the equidistance of the previous map. These new maps were then given as

H	F1	F2	F3
10	1590,6590	1345,6345	1640,7085
20	1585,6585	1345,6345	1635,7080
40	1585,6595	1345,6355	1635,7090
80	1575,6605	1345,6335	1645,7100

Table 7.2. The optimum locations for 3 facilities in Sweden. 4 square net maps of population distribution in 1965 were used as input data to the location model. The equidistance of the maps (H) assumed the values 10, 20, 40 and 80 kilometres, respectively.

input to NORLOC. The results obtained were compared with one another and with the results obtained from the original map. Stable solutions were obtained for Kronoberg county using an equidistance H equal to 1, 2, 4 and 8 km. On the other hand H equal to 16 km gave an unstable solution. In this case the map was only made up of about 50 squares, a too small a number for an analysis of this type.

The results from the sensitivity analysis indicate that the solutions obtained are only affected to a small extent by moderate variations in the size of grid square used. The model can be said to be stable in this respect. It is unfortunately impossible to prove the above claim, but it is possible to increase the probability of it being true by repeated numerical examples. In the following paragraphs three further examples are given. All of these examples use the map of population distribution in Sweden in 1965 as input data.

Table 7.2 and figure 7.2 show the location of three facilities within Sweden. It is surprising how stable the solutions produced are. The facility F_1 moves only 15 km to the west and 15 km to the north when the equidistance is raised from 10 km to 80 km. The other two facilities move even less when H is increased. That the location of a facility can be given to a greater accuracy than the equidistance of the map would probably be denied on an intuitive basis. This attitude is shown to be untrue by a study of the results shown in table 7.2. It would appear that the location of the facilities is given to an accuracy of about $0.2 H$.

The umlands are also stable when the equidistance is increased from 10 km to 80 km, a state of affairs one would expect due to the stable locations assumed by the three facilities. The small variations which occur in the umlands are a result of the fact that the umlands are aggregations of grid squares, which vary in size as the equidistance is increased.

Table 7.3 and figure 7.3 shows the case where the optimum locations of 5 facilities are chosen. Even in this case the results are very stable. The

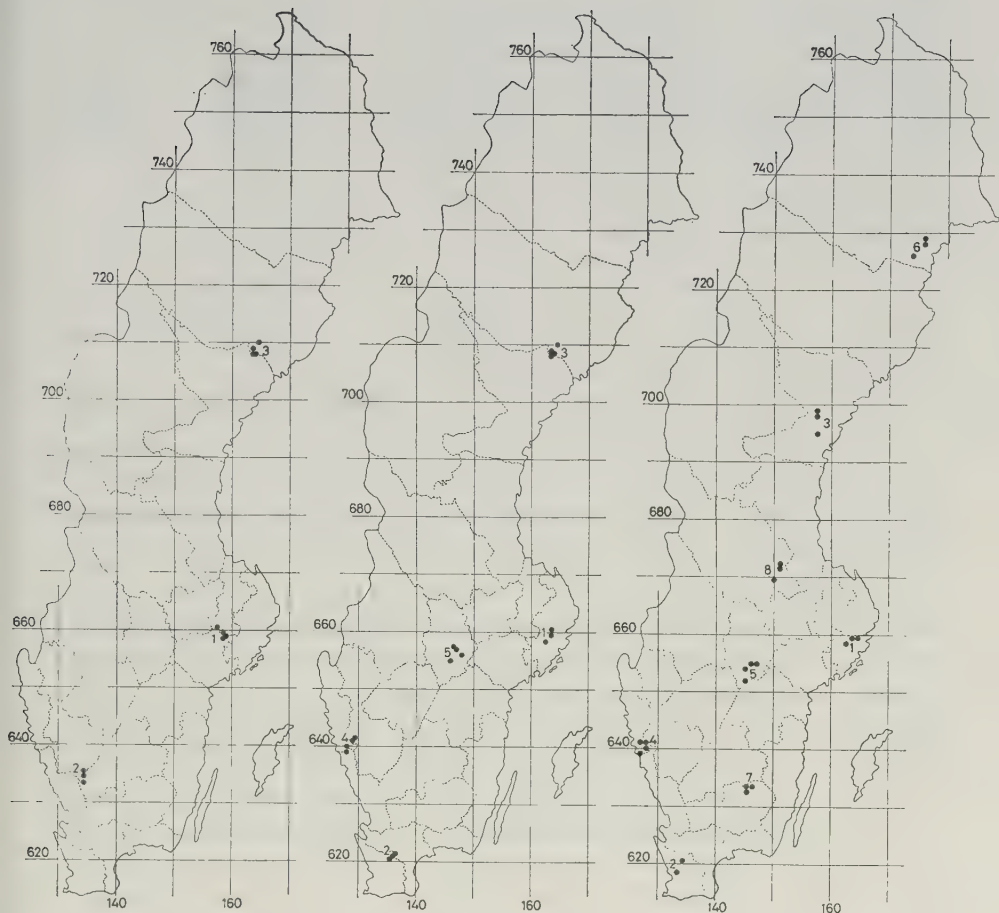


Fig. 7.2.

Fig. 7.3.

Fig. 7.4.

Figure 7.2. The optimum locations for 3 facilities in Sweden in 1965. The input data was in the form of 4 square net maps having equidistances of 10, 20, 40 and 80 km respectively.

Figure 7.3. The optimum locations for 5 facilities in Sweden in 1965. The input data was the same as that used to obtain figure 7.2.

Figure 7.4. The optimum locations for 8 facilities in Sweden in 1965. The input data was the same as that used to obtain figure 7.2.

errors occurring are still approximately $0.2 H$. The location of the umlands is also stable as a result of the stable locations assumed by the facilities.

The impression obtained, that the location model and the solutions it produces are stable when H is varied within reasonable limits, is amplified by study of figure 7.4 and table 7.4. This example shows the optimum location

H	F1	F2	F3	F4	F5
10	1625,6585	1360,6210	1640,7085	1295,6415	1465,6575
20	1625,6585	1365,6215	1635,7080	1290,6410	1470,6570
40	1635,6595	1355,6205	1645,7100	1280,6400	1480,6560
80	1635,6605	1365,6215	1635,7090	1280,6390	1460,6550

Table 7.3. The optimum locations for 5 facilities, obtained using the same input data as that used for table 7.2.

of 8 facilities, with H varying between 10 km and 80 km. Even in this example the errors are of the order of $0.2 H$.

The result from the sensitivity analysis, with respect to reasonable variations in the size of the grid squares used in the input data, shows that the model is stable in this respect. This result agrees with that obtained by Gould and others (1971). Hence, the examples described here add weight to the claim that the location model is insensitive to reasonable variations in the grid size used.

The transportation distance from the customer to the facility is increased for certain customers and decreased for others when the size of the grid square is increased. If the population distribution is such that the location function has more than one minimum then the altered transportation distances can make a local minimum into an absolute minimum and vice versa. In cases when a map with large grid squares is used the iterations can be changed so that the model fastens at a different minimum point, i.e. it gives different locations for the facilities to those obtained using a map with smaller grid squares. This phenomenon can create the impression that the model is unstable with respect to variations in the size of the grid square, when it is not actually the case. The instability is a result of the distribution of population and the local minimum points that the distribution generates and not a result of the variations in the size of the grid square.

7.3 The transportation cost function and the stability of the location model

The accuracy with which the transportation cost function must be given when using NORLOC is an important problem which requires investigation. As usual the term transportation cost has a very wide meaning. It can be the cost incurred by an individual when visiting the facility nearest to his home; or, alternatively, the time required to visit the facility, and so on.

H	F1	F2	F3	F4
10	1625,6585	1335,6185	1575,6980	1280,6410
20	1625,6585	1335,6185	1575,6980	1270,6410
40	1635,6595	1335,6185	1575,6990	1280,6400
80	1645,6595	1345,6205	1575,6950	1270,6390

H	F5	F6	F7	F8
10	1450,6540	1760,7290	1455,6325	1510,6715
20	1460,6550	1760,7290	1455,6335	1510,6715
40	1470,6550	1760,7280	1455,6335	1510,6725
80	1450,6520	1740,7260	1465,6335	1500,6695

Table 7.4. The optimum locations for 8 facilities in Sweden in 1965, as given by the location algorithm. The input data was the same as that used to obtain table 7.2.

Transportation costs can be given in table form, or in the form of a function. As a table of the type used to give transportation costs can usually be described by a function, only cases where a function is used are dealt with here. Such functions contain two parts, a fixed loading and unloading cost (terminal cost), and a pure transportation cost that varies in accordance with the distance travelled. The simplest transportation cost function is linear, the variable part of the function varies in direct proportion to the distance travelled. Other versions of the function give a reduction in the price per kilometre the longer the distance travelled. The tariff employed by the Swedish Railways is of this type. A third type of transportation cost function is that where an increase in the distance travelled leads to an increase in the cost per kilometer. Examples of this type of function include the cost of transportation by lorry, taking into account the costs incurred by rest periods and overnight stays. A similar phenomenon occurs when a traveller is forced to change from a cheaper to a more expensive transport method during a journey, for example walking, underground, bus, and airplane.

A further transportation cost function, which occurs often in practical situations, is the logistic transportation cost function. This function, when illustrated by a curve in a coordinate system, reminds one of the logistical growth curve. At the beginning there is a small increase in transportation cost as the distance increases. When a critical distance is reached a major increase in the transportation cost takes place. A further increase in the distance travelled leads once more to a relatively moderate increase in the cost of transportation. An example of such a function in operation is provided by the costs incurred by the community when transporting children to school. Over short distances the children walk or ride a bicycle to

school and the costs incurred are therefore low. When the child's home lies more than a certain critical distance away from the school the children must be allowed to use some form of public transport, bus for example. When this happens transportation costs are increased considerably. Once the child is on a bus it can travel a long distance before a further increase in transportation cost is incurred. The assumption made, in this case, is that the public transport system throughout the schools area applies a fixed price system. The above example can be described using a logistic function, even when special school buses must be used for certain pupils.

7.3.1 Power functions

The three different types of function given earlier can be described using a single function. This general function is given as formula 7.1. The value of the function $COST(D)$ is the cost incurred when transporting the customer over distance D . A refers to the terminal cost, which is taken to include the cost of loading and unloading. The parameter A can also describe delivery charges. A is equally large wherever the facility are located. This fact makes possible the exclusion of A from the location calculations. The terminal cost does not need to be given even approximately if the optimum locations are all that is required. The parameter C is used to adjust the results given by the function in order to give them in the correct units. C is therefore a scale factor, in cases where the variable part of the trans-

$$COST(D) = A + C \cdot D^B \quad (7.1)$$

portation cost function is linear C refers to the cost incurred when transporting goods or customers 1 unit of length, for example 1 km, always assuming that the distance D is given in the same units. B is an exponent which determines the way the function is to behave. If B is equal to 1 the function is linear. If B is less than 1, $COST(D)$ becomes concave, B larger than 1 gives a convex function. Figure 7.5 illustrates this effect; the figure shows a concave curve, ($B = 0.6$), a linear curve, ($B = 1$) and a convex curve, ($B = 1.6$).

Two of the parameters included in the transportation cost function $COST(D)$ affect the results obtained when the function is used in conjunction with the location model and the NORLOC program. These are B and C . A sensitivity analysis of the model with regard to the function $COST(D)$ must therefore pay regard to both B and C .

7.3.2 Parameter C

As was discussed in section 7.3.1 parameter C can be treated as a scale factor, used to ensure that the results given by the function are in the right

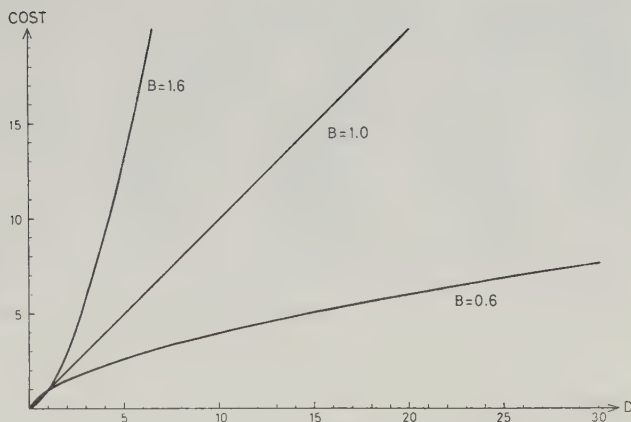


Figure 7.5. The exponent B in formula 7.1. can be given different values, which result in different curves, concave ($B = 0.6$), linear ($B = 1.0$) and convex ($B = 1.6$).

units. The value given to C depends to a great degree on the units used for D . A further factor which affects the choice of a value for C is the cost of the transportation method used. C is small for a walker but much larger for a taxi passenger.

It can be shown that the parameter C does not effect the locations given to facilities by the location model by the setting in of formula 7.1 in formula 5.5 which describes the location function. One can then see that C can be placed before the summation symbols to give formula 7.2. The value $D_{ij,Fk}$ refers to the distance between the grid square (i,j) and the facility Fk .

$$f(X_{F1}, \dots, Y_{FN}) = C \sum_{i=1}^L \sum_{j=1}^W P_{ij} \min_{1 \leq k \leq N} (D_{ij,Fk}^B) \quad (7.2)$$

It can be clearly seen that C is a scale factor, thus it can be given any value larger than 0 when the location calculations are carried out without the choice of locations being affected. A good choice for this value would be 1. If at a later stage one is interested in the actual value of $f(X_{F1}, \dots, Y_{FN})$ it can be obtained by multiplying the value available by the actual value of C .

The allocation of the value 0 to C creates a location model that is insensitive to distance; the only costs studied are the terminal costs. This version of the model is highly unrealistic as the number of facilities located and the positions chosen for them has no effect on the result.

7.3.3 Exponent B

The exponent B which forms part of the transportation cost function $COST(D) = A + CD^B$ controls how sensitive the location model is to varia-

Expo- nent	P o s i t i o n s		
	F1	F2	F3
0.6	1625,6585	1325,6385	1625,7040
0.8	1615,6585	1340,6365	1630,7065
1.0	1590,6590	1345,6345	1640,7085
1.2	1570,6595	1345,6340	1645,7110
1.4	1565,6595	1350,6335	1645,7115
1.6	1555,6595	1350,6330	1645,7125

Table 7.5. The effect of different values of B on the optimum locations chosen for 3 facilities. Data: Square net map of population distribution in Sweden in 1965.

tions in distance. The smaller the value assigned to B the less the importance attached to longer distances. If B is equal to 0 the transportation function ceases to react to changes in the distance travelled. $COST(D)$ is then equal to $A + C$ for all values of D . This version of the function is of limited interest due to the fact that no optimum location can be determined since the value of the function $f(X_{F1}, \dots, Y_{FN})$ is equal to $P \cdot (A + C)$ for all locations of $F1, \dots, FN$. P is equal to the total number of customers using facilities $F1, \dots, FN$.

Chapter 6 dealt with an example where there existed both an absolute and a local minimum, when locating two facilities in the area in question, see figure 6.3. In the case described in chapter 6 several versions of the transportation cost function were used. In these different versions the exponent B varied between 0.9, 1.0 and 1.3. The resulting functions were concave, linear and convex. The result with all three versions showed an absolute and a local minimum. The locations assumed by the facilities varied very little as a result of the functions used, see table 6.1 and figures 6.6–6.8. It can therefore be said that chapter 6 contains an attempt at a sensitivity analysis of the model in respect of different values of parameter B .

In the sensitivity analysis described in this section the input data has been obtained from the square net map of population distribution in Sweden for the year 1965, (figures 4.5 and 4.6). The actual analysis is begun by the making of a decision about the number of facilities it is required to locate. In this example the number of facilities N is taken to be equal to 3, 6 and 12. A decision is also required concerning the values to be used for B . In this case the values chosen were 0.6, 0.8, 1.0, 1.2, 1.4, and 1.6. Taken together these values of N and B result in 18 different examples.

Table 7.5 shows the optimum location of three facilities for different values of B . It can be seen from the table that the facilities move more and more to the centres of their respective umlands, as the value of B rises. $F1$, which

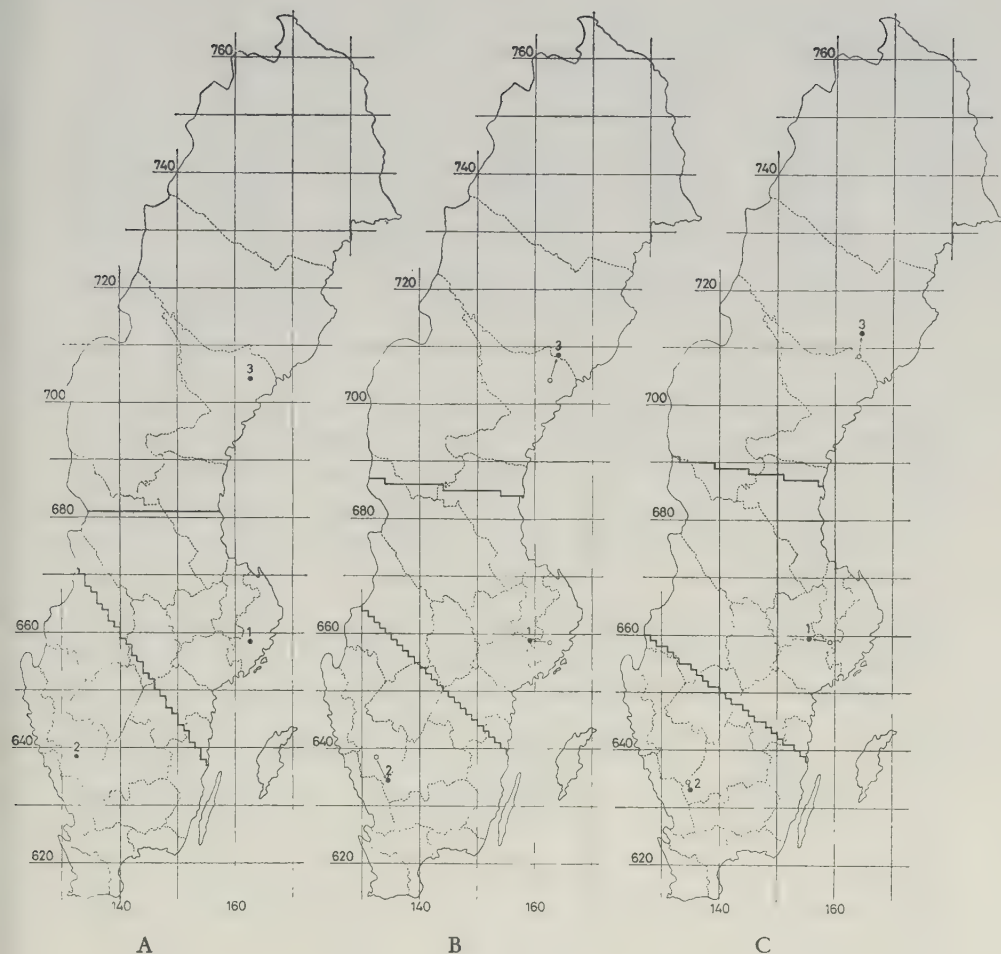


Figure 7.6. The optimum locations for 3 facilities using 3 different values of B in the transport cost function. Fig. A: $B = 0.6$, fig. B: $B = 1.0$ and fig. C: $B = 1.6$.

in the first instance is located in the vicinity of Stockholm, when the value of B is 0.6, moves 5—25 km westwards every time the value of B is increased. The sum of these movements is 70 km westwards. At the same time as F_1 moves even F_2 and F_3 move roughly the same distance. The direction taken by F_2 is mainly southwards, that taken by F_3 mainly northwards, even if both of these facilities move somewhat to the east.

Figure 7.6 shows that as a result of the movements of F_1 , F_2 and F_3 the umland of F_1 is extended in both a northerly and a southerly direction. The maps in the figure show the boundaries assumed by the different

Expo- nent	P o s i t i o n s		F3	F4	F5	F6
	F1	F2				
0.6	1625,6585	1345,6195	1575,6925	1275,6405	1465,6575	1765,7255
0.8	1625,6585	1355,6205	1570,6930	1275,6405	1465,6575	1755,7265
1.0	1625,6585	1360,6210	1560,6930	1295,6415	1465,6575	1750,7260
1.2	1625,6585	1365,6215	1555,6935	1300,6415	1465,6575	1745,7260
1.4	1620,6585	1370,6215	1550,6935	1310,6415	1465,6575	1740,7260
1.6	1615,6590	1375,6220	1545,6940	1310,6420	1465,6575	1740,7265

Table 7.6. The effect of different values of B on the optimum locations chosen for 6 facilities. Data: Square net map of population distribution in Sweden in 1965.

umlands in the cases when B is equal to 0.6, 1.0 and 1.6. In cases such as those where B is equal to 0.8 the umlands' boundaries are not shown, but can be assumed to lie between the boundaries shown in figure 7.6 A and 7.6 B. The same is true for values of B lying between 1.0 and 1.6. In such cases the boundary can be assumed to lie between those shown in figure 7.6 B and 7.6 C.

A study of the different maps included in figure 7.6 shows that each facility has approximately the same umland irrespective of which value of B was used. In all cases a district is obtained in the southwest of Sweden, including Småland, Västergötland and Dalsland. In a similar way the boundary between northern Sweden and central Sweden follows the southern boundaries of Härjedalen and Medelpad. The names and locations of the provinces in Sweden are given in figure 7.16 on page 246.

The major part of the county of Värmland is shown in figure 7.6 A as being a part of the southwestern region. This part of Värmland should in fact form part of the central Swedish region. Lake Vänern lies just to the south of Värmland and cuts off any communication with southwestern Sweden in a very effective manner. If Lake Vänern had been described to the program as a barrier, with all distances calculated around the lake instead of across it, then NORLOC would have included Värmland in the central Swedish region. It would be fairly straightforward to extend NORLOC with a barrier program, but at the cost of a much slower program. It is a very much cheaper procedure to manually adjust the boundaries of the umlands, so as to coincide with existing natural, administrative or other boundaries.

In the case where N is equal to 3 the umlands one obtains are the same, irrespective of the value of B . It is therefore possible to use a somewhat incorrect value of B and still obtain a correct solution. The location model is therefore relatively stable with respect to B . One should also bear in mind that it is possible to divide Sweden into 4 natural areas, on the basis

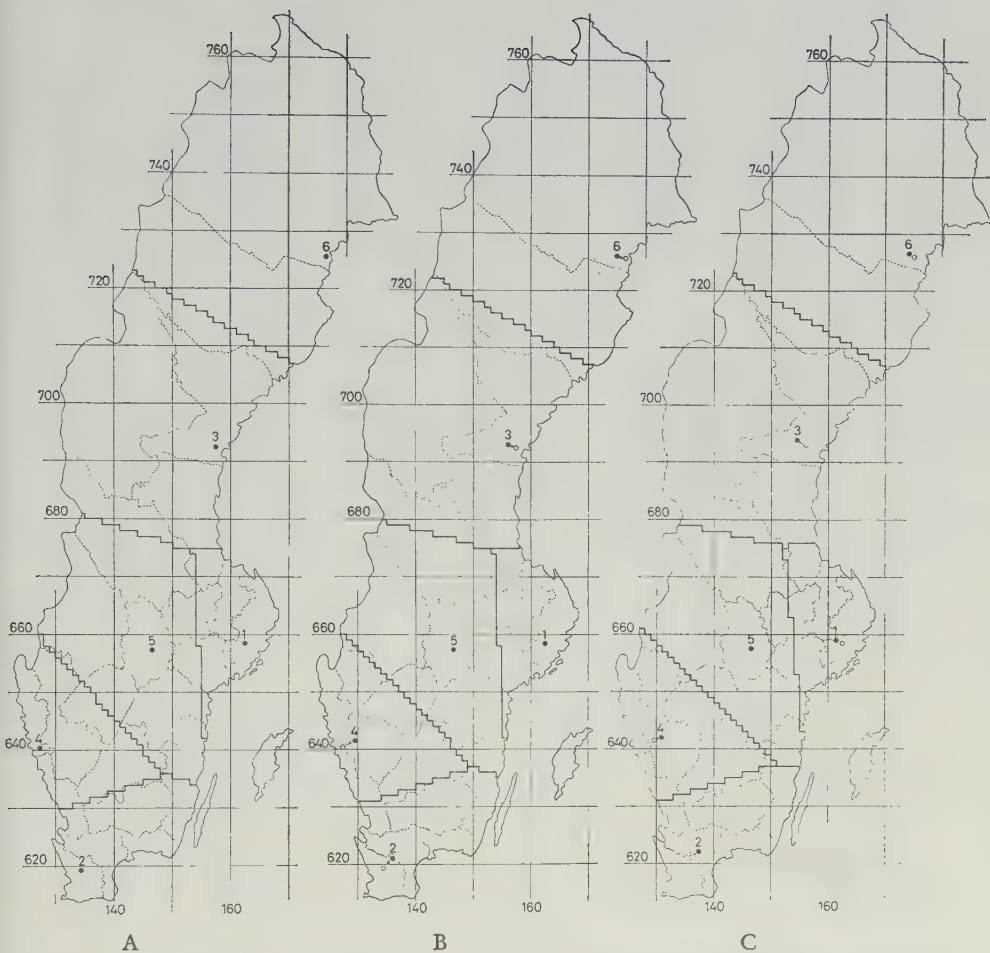


Figure 7.7. The optimum locations for 6 facilities obtained when using different values of B in the transport cost function. The data used was for 1965. Fig. A: $B = 0.6$, fig. B: $B = 1.0$ and fig. C: $B = 1.6$.

of population distribution, these are north Sweden and the three major urban conurbations. This fact makes the chances that $f(X_{F1}, \dots, Y_{F3})$ has one or several local minimum values quite large. The facility $F2$ must serve both the Göteborg and Malmö regions. Changes in the value of B cause $F2$ to move towards one or the other of these conurbations. It is therefore a matter of surprise that the changes in the position of $F2$ are as small as they are when B varies between 0.6 and 1.6.

An extension of the sensitivity analysis to include the location of 6 facilities in Sweden reinforces the impressions one obtains when working with only

Expo- nent	P o s i t i o n s		F3	F4	F5	F6
	F1	F2				
0.6	1625,6585	1465,6265	1605,6985	1275,6405	1535,6605	1775,7295
0.8	1625,6585	1465,6270	1605,6985	1275,6405	1520,6600	1770,7295
1.0	1625,6585	1470,6275	1610,6990	1275,6405	1515,6595	1765,7295
1.2	1625,6585	1475,6280	1620,7005	1275,6405	1515,6595	1760,7300
1.4	1625,6585	1475,6280	1625,7010	1280,6410	1510,6600	1755,7305
1.6	1625,6585	1470,6280	1625,7010	1280,6410	1510,6600	1750,7305

	F7	F8	F9	F10	F11	F12
0.6	1335,6175	1365,6585	1405,6405	1495,6475	1505,6725	1445,7005
0.8	1330,6175	1365,6585	1405,6405	1495,6475	1525,6735	1445,7005
1.0	1335,6185	1365,6585	1405,6405	1510,6485	1520,6730	1445,7005
1.2	1335,6190	1365,6585	1400,6415	1515,6480	1520,6740	1445,7005
1.4	1335,6190	1365,6590	1400,6415	1515,6480	1525,6745	1450,7005
1.6	1340,6195	1365,6590	1395,6425	1520,6475	1525,6750	1450,7000

Table 7.7. The effect of different values of the exponent B in the transport cost function $COST = A + CDB$ used by the location algorithm when determining the optimum location of 12 facilities, using data describing the distribution of population in 1965. The initial values are those obtained when B is given the value 1.6, i.e. the locations shown in the bottom row of the table.

three facilities. Table 7.6 and figure 7.7 show that the facilities move nearer and nearer the geometric centre of their respective umland, the greater the value allocated to B . In this case facilities $F1$ and $F5$ remain in approximately the same place, whereas the other 4 move from an approximately coastal position to a position inland. An inspection of the umlands and the form they assume with different values of B shows that, in this respect, the results obtained are very stable (figure 7.7). The umlands obtained are practically the same, irrespective of the values assigned to B .

Table 7.7 shows how variations in the exponent B forming part of the transportation cost function influence the locations chosen for 12 facilities by the location program. Several of the facilities are located in the same or exactly the same position, irrespective of the value of B . $F11$ is the facility most affected by changes in the value of B . This facility moves 25 km to the north and 20 km to the west when the value of B is increased from 0.6 to 1.6. As usual, a certain amount of movement of facilities towards the centres of their respective umlands occurs when the value of B is increased. Even so, the actual umlands are relatively unaffected, see figure 7.8. The solutions given by the location model are stable even in respect to B , with notice taken of the limitations mentioned earlier.

In the case where $N=12$ an interesting situation arises, due to the fact that the function $f(X_{F1}, \dots, Y_{F12})$ has two minimum points, also a certain

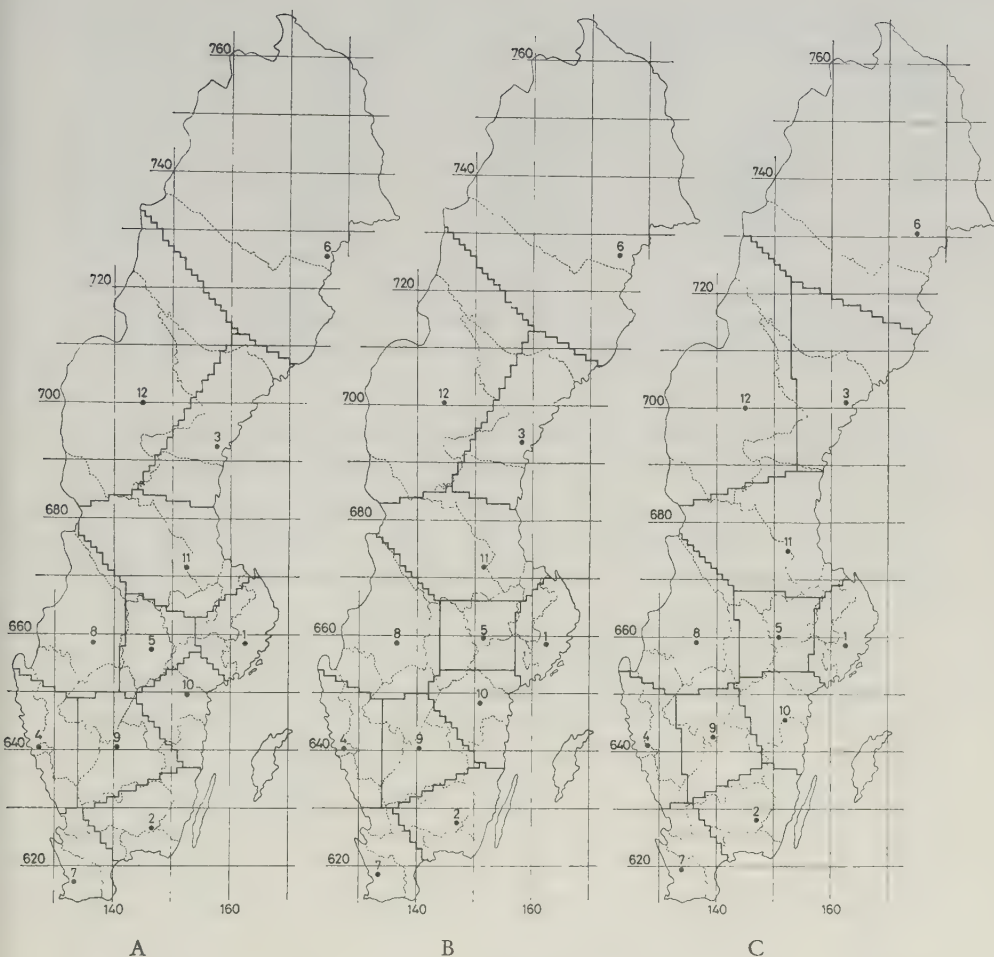


Figure 7.8. The optimum locations for 12 facilities, using different values of B in the transport cost function. The data used was for 1965. Fig. A: $B = 0.6$, fig. B: $B = 1.0$ and fig. C: $B = 1.6$.

amount of interference between the two solutions arose. In the first case an arrangement was generated with B equal to 0.6, an increase in the value of B to 0.8 produced a further different arrangement. A further increase in the value of B to 1.0 produced a return to the first arrangement, and so on. This was a great surprise, as it meant that the arrangements obtained were not stable with respect to the exponent B in the transportation cost function. A continued analysis lead to the conclusion that it was a case of two different solutions. This conclusion was drawn from the results obtained when using the first arrangement, with $B = 0.6$, as the initial

Expo- nent	P o s i t i o n s			
	F3	F5	F6	F10
0.6	1575,6925	1465,6575	1765,7255	1525,6495
0.8	1575,6925	1465,6575	1755,7265	1515,6490
1.0	1580,6935	1515,6595	1750,7265	1510,6485
1.2	1580,6940	1515,6595	1750,7265	1515,6480
1.4	1620,7005	1510,6595	1755,7305	1515,6475
1.6	1625,7010	1510,6600	1750,7305	1520,6475

Table 7.8. The table illustrates a local minimum to the function $f(X_{F1}, \dots, Y_{F12})$. $F3$, $F5$, $F6$ and $F10$ have different locations to those given in table 7.7, the remaining facilities acquired identical or nearly identical locations to those given in table 7.7.

arrangement for the following executions of the location program. It can be seen from table 7.8 that it was primarily facilities $F3$, $F5$, $F6$ and $F10$ which moved, respective to the locations given in table 7.7. $B=1.6$ gave a return to the original arrangement. The allocation of the value 0.6 to the exponent B leads to two clearly differing arrangements, as does the use of 0.8 as the value of B .

An inspection of table 7.9 shows that the two minimum values are very close to one another, especially in the cases where B is 0.6 and B is 1.6. This phenomenon shows that care must be taken to avoid missing solutions which are as good, or nearly as good, as those the model gives.

In the examples described in this section the map of population distribution in Sweden dated 1965 was used as the input data, in conjunction with varying values of the exponent B . B was allowed to vary between 0.6 and 1.6, a variation over a wide range in the case of an exponent. Figure 7.9 shows 15 different maps of Skåne. NORLOC has in these 15 cases been used to locate 5 facilities within the area in question, with the value of B varying between 0.5 and 5.0. Thus, in this case the exponent B varies

Exponent	Total transportation costs		Relative number
	Min 1	Min 2	
0.6	$7.0787 \cdot 10^7$	$7.0793 \cdot 10^7$	100.009
0.8	$16.075 \cdot 10^7$	$16.201 \cdot 10^7$	100.78
1.0	$37.080 \cdot 10^7$	$37.226 \cdot 10^7$	100.39
1.2	$86.833 \cdot 10^7$	$87.609 \cdot 10^7$	100.89
1.4	$205.66 \cdot 10^7$	$205.71 \cdot 10^7$	100.024
1.6	$492.54 \cdot 10^7$	$492.55 \cdot 10^7$	100.002

Table 7.9. The total transport cost incurred when the location model is used to chose the locations for 12 facilities. It can be seen, from the right hand column, that the difference between the values obtained at the two minimum points is so small as to be unimportant.

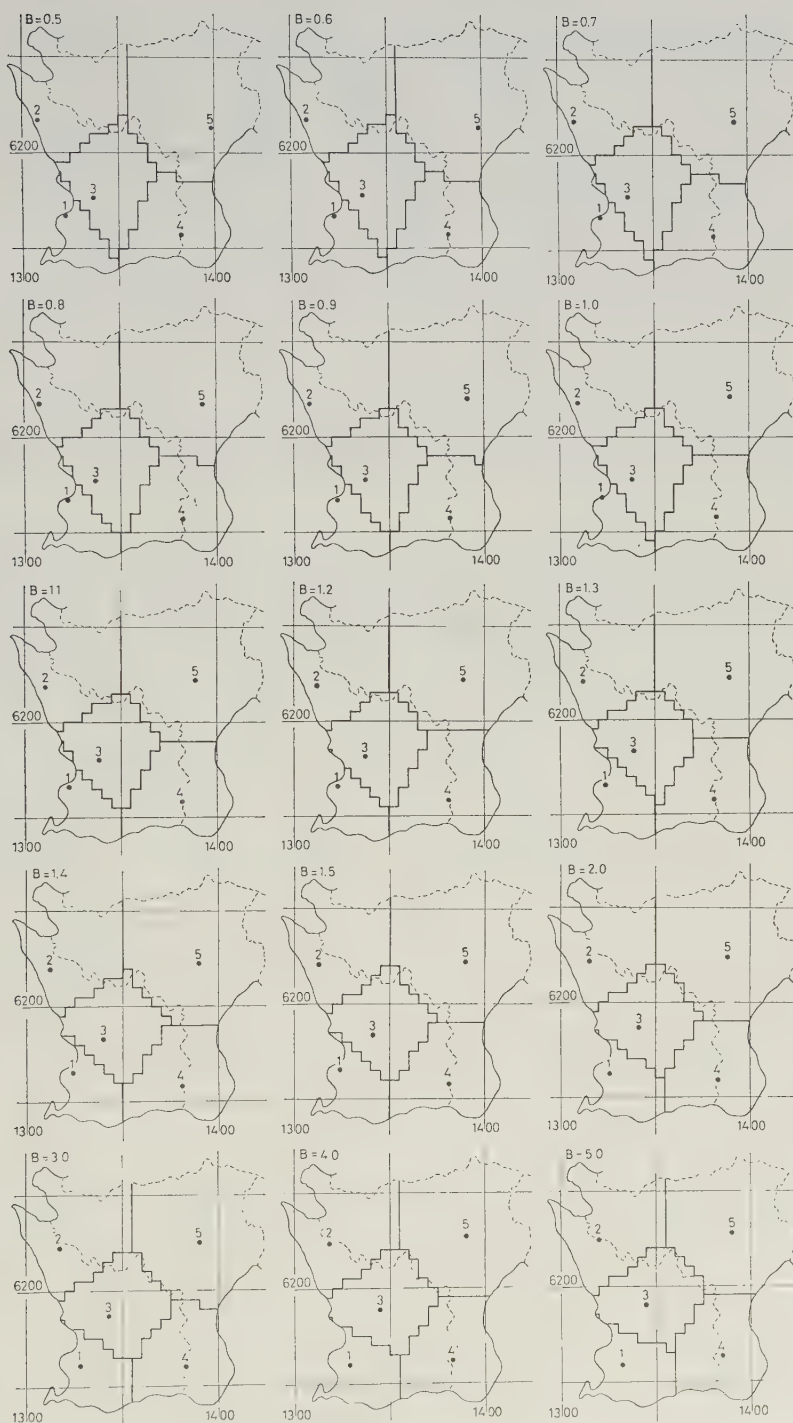


Figure 7.9. The optimum location of 5 facilities in Skåne in 1965. The exponent B in the transport cost function was varied between 0.5 and 5.0.

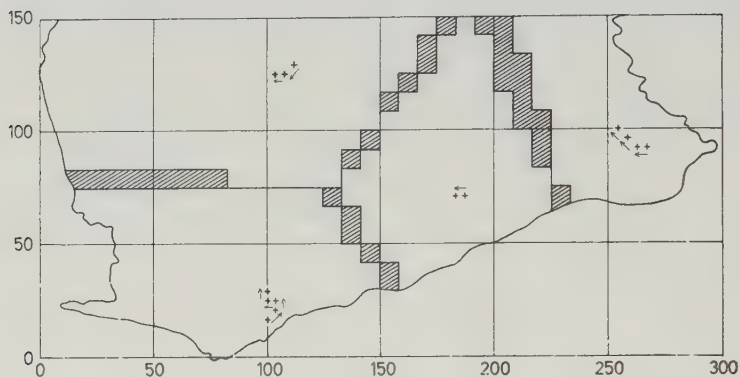


Figure 7.10. The optimum locations for 4 facilities in southern Ghana. The exponent B was varied between 0.5 and 1.5.

to an even greater extent than in the previous examples. It can be seen from the figure, that the arrangements obtained are highly stable. The locations allocated to the different facilities are only to a very small extent dependent on the value of B . An increase in the value of B moves the facilities nearer to the geometric centre of their respective umlands.

Figure 7.10 and table 7.10 illustrate a further example showing that the location model is stable with respect to the value of B in the transportation cost function. This example uses a square net map of population distribution in southern Ghana (Claeson, Godlund & Örtenblad 1969). The value of B varies between 0.5 and 1.5. These variations in B caused small changes in the location of the 4 facilities. The facility $F3$ only moves a half unit to the west, facilities $F1$, $F2$ and $F4$ move one unit in the same direction, these variations occurring when B increases from 0.5 to 1.5. Movement in a north—south direction is even less evident. It can be seen from figure 7.10 that the umlands are highly stable. Changes in the location of facilities and their umlands are unimportant and can be accommodated in the rounding off errors occurring in this type of calculation.

One can summarize the above by saying that the location model is relatively insensitive to variations in exponent B in the function $COST(D) = A + CD^B$. It is therefore not necessary to give the value of B to any great accuracy. In conjunction with this statement it should also be pointed out that the distribution of population can be such that more than one minimum exists. In such situations an alteration in the value of B can transform a local minimum to an absolute minimum. Interference can also occur between different minimum solutions, thus giving the impression that the solution is unstable. However, the reason for this apparent instability is

Expo- nent	P o s i t i o n s			
	F1	F2	F3	F4
0.7	13.5,15.5	12.5,2.5	22.5,8.5	31.5,11.0
0.9	13.0,15.0	12.5,3.0	22.5,8.5	31.0,11.5
1.1	13.0,15.0	12.0,3.0	22.5,8.5	31.0,11.5
1.3	13.9,15.0	12.0,3.5	22.5,8.5	31.0,11.5
1.5	12.5,15.0	12.0,3.5	22.0,8.5	30.5,12.0

Table 7.10. The optimum locations for 4 facilities in southern Ghana, obtained using the location algorithm. The exponent B varies between 0.7 and 1.5.

to be found in the spatial distribution of the customers and not in the sensitivity of the model to variations in the exponent B .

7.3.4 The logistic variant of the transportation cost function

The logistic growth function is characterised by the fact that a very large increase in the value of the function occurs when the independent variable approaches and passes a certain value. On either side of this value a change in the value of the independent variable causes only a moderate increase in the value of the dependent variable. The logistic function can only vary between 0 and l in value, thus prohibiting its application as transportation cost function. Fortunately it is relatively simple to modify formula 7.1 in such a way as to give an approximation of the logistic function. This new function is shown as formula 7.3 below. In this formula A , B , C and D refer to the same variables as in formula 7.1. E is the value lying at the

$$COST(D) = A + C \cdot \text{sign}(D - E) \cdot (\text{abs}(D - E))^B + CE^B \quad (7.3)$$

centre of the interval within which the large increase in the value of the dependent variable occurs. The formula includes two mathematical terms, $\text{sign}(D - E)$, which refers to the sign of the difference between D and E , and $\text{abs}(D - E)$, which refers to the absolute value of this difference. The curve of function 7.3, in the case when B is equal to $1/3$, is shown in figure 7.11. From this figure it can be seen that the function behaves in the manner required. A drastic rise in transportation cost occurs in the interval immediately on either side of E , otherwise an increase in the value of D results in a very moderate increase in transportation cost.

Allocation of the value 0 to D makes $D - E$ less than 0 and $\text{sign}(D - E)$ becomes -1 . This is true if E is larger than 0. The term $\text{abs}(D - E)$ is, in this case, equal to E . Substitution in 7.3 above gives $COST(D)$ as being equal to $A - CE^B + CE^B$, which is equal to A . Allocation of the value 0 to E makes $\text{sign}(D - E)$ equal to $+1$, at the same time as $\text{abs}(D - E)$ becomes

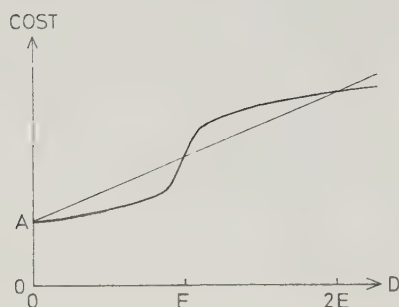


Figure 7.11. The curve of the logistic version of the transport cost function with exponent $B = 1/3$.

equal to D . Formula 7.3 can be simplified in this case to $COST(D) = A + CD^B$, this being the same formula as formula 7.1. It would have been quite satisfactory if only formula 7.3 had been included in the NORLOC program. Due to the fact that less computer time is required when calculating transportation cost using formula 7.1 both formulas 7.1 and 7.3 are present in the program, together with the linear function $COST(D) = A + CD$.

It has been shown earlier that the model allocates approximately the same positions and umlands to facilities, whichever version of the transportation cost function is used. It was felt necessary to extend the sensitivity analysis to include studies of the extent to which use of the logistic transportation cost function effects the stability of the model's solutions. Southern Sweden is a suitable area for such a study. The number of facilities chosen was 7, in order to allow comparison with the example where a linear function was used. The values chosen for the parameters in the logistic function were $B = 1/3$, $C = 8.55$ and $E = 25$.

The results obtained are shown in table 7.11 and figure 7.12. The input data used was a square net map with an equidistance of 5 km, the map was constructed using coordinate based real estate data. It can be seen from

B	C	E	F1	F2	F3	F4
1	1.00	0	1325,6167	1310,6217	1322,6285	1397,6212
1/3	8.55	25	1330,6165	1315,6220	1322,6280	1395,6220

B	C	E	F5	F6	F7
1	1.00	0	1435,6305	1527,6287	1477,6232
1/3	8.55	25	1435,6305	1527,6290	1480,6235

Table 7.11. The optimum locations for 7 facilities in southern Sweden when using a linear ($B = 1.0$, $E = 0$) and a logistic ($B = 1/3$, $E = 25$) transport cost function.



Figure 7.12. The optimum locations for 7 facilities in southern Sweden. Both a linear and a logistic transport cost function were used. The locations are given by means of solid and open circles, respectively.

figure 7.12 that one obtains the same position and umland for the facilities, irrespective of which function is used, logistic or linear. It can therefore be said that even in this case the model gives stable solutions.

7.4 Sensitivity to the use of straight line distance instead of real road distance

The location model can be criticized for its use of straight line distance, i.e. distance measured as the crow flies, instead of the real road distance. The majority of persons feel, intuitively, that the road network has an important effect on choice of location. A sensitivity analysis of the model, with respect to the way distance is measured, is therefore called for in order to see to what degree the stability of the chosen locations is affected. In order to do this it is necessary to define the term real road distance.

7.4.1 Shortest routes and the location model

The real road distance between two points can be defined as the length of the route, via the road network, which a person uses when travelling between the two points. This definition is not complete, due to the fact

that the choice of route depends on the individual person. The way individuals choose routes is not fully understood and it is not therefore possible to define what type of route the majority of individuals choose. If a rational individual has access to all information about the road network he would probably choose an optimum route between the two points. An optimum route can be defined in a number of ways. The shortest route, the quickest route, the safest route, the most beautiful route and the cheapest route are all optimal to a certain extent. The majority of persons feel that the difference between the optimum and the chosen route is relatively small, although investigations have shown considerable differences between the two (Nordbeck & Rystedt 1969).

The difference between a chosen route and an optimum route is fairly small compared with the difference between the chosen route and the straight line distance. From a theoretical point of view the use of the shortest distance via a road network is a considerable improvement over distance measured as the crow flies. A more complicated version of the location model, capable of working with distance measured via a road network, requires access to the length of the shortest route between the majority of pairs of nodes in the network. These lengths can be measured using the shortest route program NORCAS. Unfortunately this program requires a great deal of computer storage. The size of the road network which can be handled is therefore small. In the case of the UNIVAC 1108 computer at Lund (October 1972) it is not possible to work with a network containing much more than 300 nodes.

7.4.2 Sensitivity analysis using Lund data

The road network in Lund contains about 1000 nodes and is therefore too large to be handled by NORCAS. On the other hand the main road network in Lund contains only 232 nodes, see figure 7.13. The parts of the road network not included in this network are of a local character. A route limited to the main road network is probably somewhat longer than the route via the total net, but this difference is felt to be sufficiently small to be unimportant. The shortest route via the main road network is probably nearer the route chosen by individuals in reality. This has been found to be the case with respect to the length of the route (Nordbeck & Rystedt 1969).

The main road network is described to the NORCAS program and a matrix of distances prepared. This matrix contains the length of the shortest route between every pair of nodes.

The real estate register in Lund is coordinate based (Bergström 1971) and can thus be used to prepare square net maps of population distributions.



Figure 7.13. The main road network in Lund.

These square net maps, together with the coordinate based road network, make Lund suitable for the sensitivity analysis with respect to different approaches to the measurement of distance.

The analysis was carried out in the following manner. The optimum locations of 5 high schools were chosen using data describing the distribution of pupils between 13 and 15 years of age. Every cell in the square net map shown in figure 7.14 was taken in turn and the link in the road network lying nearest to it determined. This results in two nodes being associated with every grid square. These nodes are stored in a special matrix. Using the information in this matrix it is possible to determine in which mesh of the road network each grid square lies. The distance between two grid squares in the population distribution map consists of three parts: (1) the distance from the initial square's centre to the nearest link in the road network, (2) the distance, via the road network to the nearest link to the goal square and (3) the distance from this link to the centre of the goal

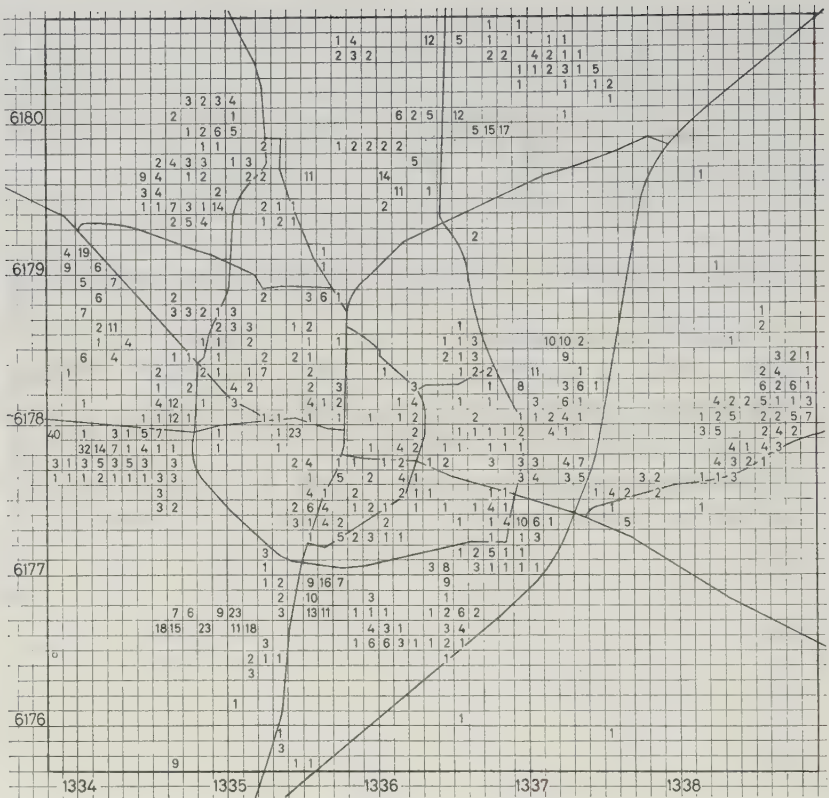


Figure 7.14. The distribution of schoolchildren in Lund in 1970 (aged between 13 and 15 years). Reference area: a square with sides 100 metres long.

square. The length of journey via the road network is already available using the distance matrix generated by NORCAS. The two other components are calculated using straight line distance multiplied by a coefficient $q = 1.2$ (Nordbeck & Rystedt 1969b). If all the nodes forming the mesh in which the origin of the journey lies are used, together with all the nodes surrounding the destination point, a large number of different routes are obtained. In order to reduce this number only the nodes lying at the ends of the nearest link are used. The shortest length of the resulting routes is then used as input data to the location model, when calculating transport costs between the two grid squares.

Figure 7.15 illustrates the optimum location of 5 high schools, with regard to the distribution of schoolchildren in the age groups concerned in 1970. Each of the 5 locations is shown by two squares; a square with a dotted perimeter in the case of locations chosen using straight line distances, and

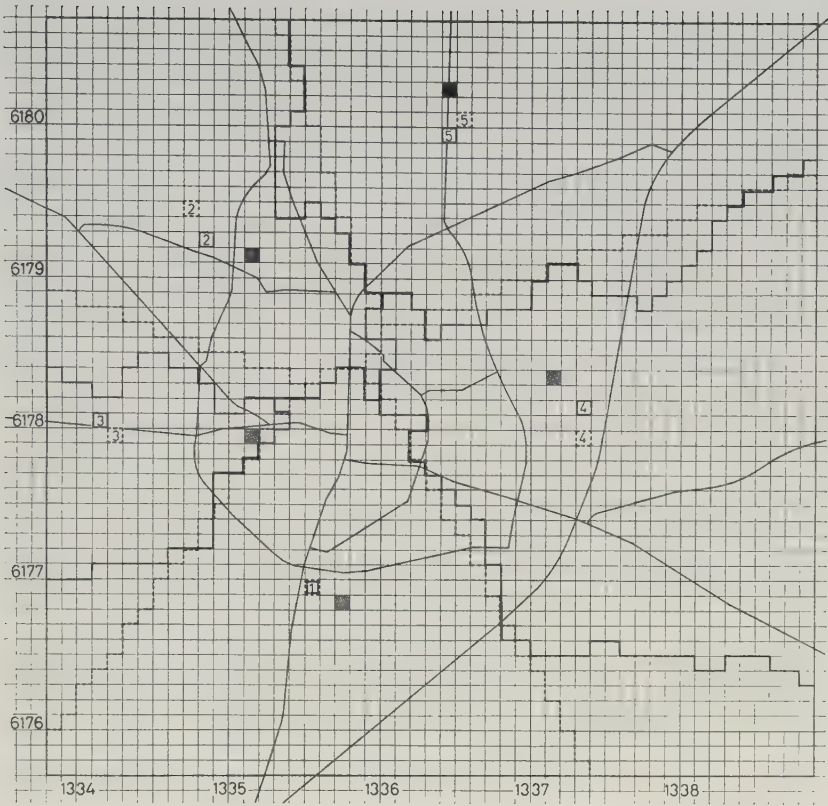


Figure 7.15. The optimum location of 5 high schools in Lund with regard to the distribution of schoolchildren in 1970 (see figure 7.13). Squares with a dotted perimeter show the locations of the schools when straight line distances are used in the model. Squares with a solid perimeter belong to the locations chosen using real road distances.

a square with solid lines in the case of locations chosen using actual distance via the road network. The boundaries between the different school districts are shown in the same way, with dotted and solid lines. The shaded squares indicate the locations of existing schools.

It can be seen from figure 7.15 that the schools are located in approximately the same positions using both approaches to the measurement of distance. School number 1 is in exactly the same location while schools number 3 and 5 move 1 grid square, to the northwest in the case of school number 3 and to the southwest in the case of school number 5. These moves occur when the more complicated approach to the measurement of distance is used. The largest difference occurs in the case of schools number 2 (300 m) and 4 (200 m). If all 5 schools are taken together it is

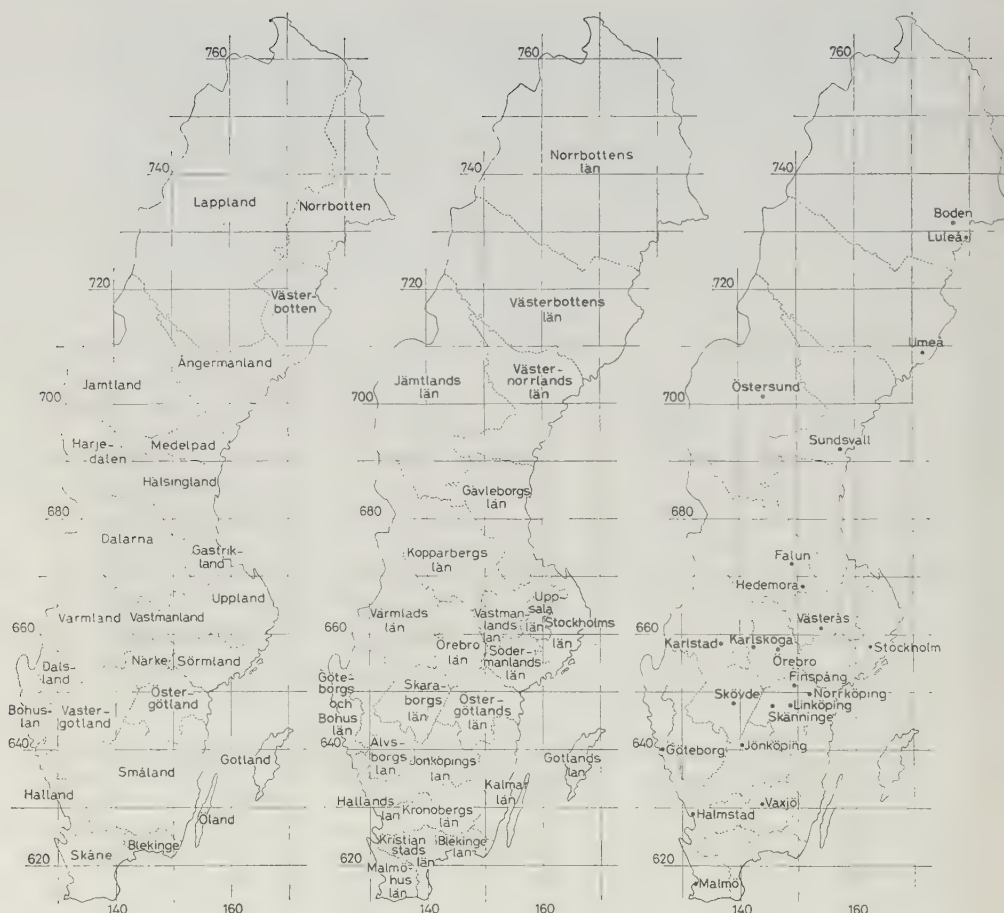


Figure 7.16. This chapter refers to a number of provinces (landskap) and counties (län) in Sweden. In order to make it easier for the reader the names of the Swedish provinces and counties are given in the left hand map and the central map, respectively. Also, some Swedish towns are referred to in chapters 7–10. The names and locations of these towns are given in the right hand map.

impossible to say which of the two sets of locations is the best. The difference between them is unimportant and the two sets of locations can for practical purposes be treated as one. The use of distance measured via a road network does not give any clear advantage over the use of straight line distance. Against this must be set the fact that use of distance measured via a network incurs computer processing costs 10 to 15 times greater than those incurred when straight line distances are used. Further costs are also incurred in preparing the description of the road network and its

transfer to computer readable form. The increased costs for the use of the more complicated model are not compensated for by increased precision in the results obtained.

A comparison of the locations produced by the computer program with the actual locations is also of interest. When school number 1 was built the Klostergården housing district in the southwest of Lund had not been built. If attention is paid to the changes in the distribution of the school pupils which have occurred since its construction one can say that its location was nearly optimal. The same is true for school number 5. Originally it was too far to the north, but the newly constructed housing districts to the north of Lund have compensated for this. School number 2 lies on the wrong side of a busy road and has a very unfortunate location. It should have been located 500—600 metres to the northwest. Its existing location is totally unacceptable, due to the severe traffic in the vicinity. School number 4 lies in the right location, according to the computer model, and its siting is fully acceptable. School number 3 lies too near the town centre. In this case one should remember that many of the pupils going to this school live outside Lund and travel by bus to school. These pupils are not included in the location calculations, although in fact they should be given a point of origin at the bus station. The school would as a result of this have been located further towards the east, but not as far to the east as it lies at present. This school is wrongly located, a fact which will be accentuated if Lund extends itself to the west.

7.5 Conclusions

The sensitivity analysis described above shows that it is possible to use a somewhat incorrect and coarse square net map as input to the location model and still obtain solutions to the location problem of sufficient accuracy. This is also the case when an approximate transport cost function is used with distances measured as straight line distances. The next chapter shows that the model is also stable with respect to changes in the distribution of population within regions. The sensitivity analysis described in this report are of the numerical proof type and are therefore, in theory, only applicable to the example in question. Even so, it is possible to say that none of the results support the contention that an accurate, large scale square net map must be used as input data or that the actual transportation costs must be known, both in terms of the function used and the way distances are measured.

8 The stability of the location model with respect to changes in the distribution of population through time

In chapter 4 a number of maps of population distribution in Sweden for the years 1855, 1917, and 1965 were presented. A number of major alterations in the distribution of the population occurred during the period 1855—1965. During the latter part of the 19th century and the beginning of the 20th a large scale colonisation of the previously unpopulated inland areas in the north of Sweden occurred. At the same time the urban population increased considerably. Stockholm increased from 98 000 to 371 000 between 1855 and 1917. The increase in population in Malmö and Göteborg was relatively greater. Göteborg increased from 29 500 in 1855 to 210 000 in 1917. In the case of Malmö the increase was 15 800 to 100 800. The increase in population for the country as a whole was from 3 700 000 to 5 600 000, a more moderate increase than that occurring in the three major towns.

The bringing under cultivation of previously unused land occurred in Sweden well into the 20th century. It was discontinued at the end of the second World War and replaced by its opposite, the cessation of agriculture in certain areas and the planting of forests. From the middle of the 1930's up to today the movement of population has been to the towns, with a thinning out of the rural population as a result (Dahlberg & Ödmann 1969). The development of Sweden into an urbanised society has only been happening on a large scale from 1935.

8.1 Changes in the distribution of population occurring within regions and between regions

Changes in the distribution of population can occur within regions or between regions. A change occurring between regions means that one

region gains population at the expense of another. Such changes have taken place in mining districts, where population was previously very small, the increase in population taking place when new deposits are discovered and exploited. This type of alteration leads to the location model giving two different sets of results depending on whether the data describes the situation before or after the changes in population distribution took place. The second type of alteration in population distribution occurs when population distribution within a region alters. It occurs e.g. when residents of rural districts move into the neighbouring towns. The population of the different regions are relatively the same throughout this type of population movement. The numbers shown in the squares of the maps can differ, but the distribution remains the same. In such cases the location model gives stable solutions, irrespective of the point in time the input data describes.

It is, of course, only in special cases that changes in population distribution take place purely between or within regions. In most cases a combination of both occurs. The alterations in population distribution can, therefore, in most cases only be classified into mainly within the region or mainly between regions. This classification can be done using correlation analysis. If the changes have taken place within the regions the two maps will show a high coefficient of correlation. On the other hand, changes occurring between regions make the correlation coefficient smaller or even negative. A correlation analysis of the above type was carried out using the maps shown in figures 4.1 to 4.6. The results show that the majority of changes have occurred within regions. The correlation coefficients obtained were 0.85 for the period between 1855 and 1917, 0.90 for the period between 1917 and 1965 and 0.68 for the period between 1855 and 1965. These coefficients were not obtained using the values shown in the grid squares of the different maps in chapter 4. Instead 16 of the grid squares were combined to create grid squares with sides 40 km long. The new large grid squares overlapped one another, using an overlapping constant of 2. The overlapping resulted in about 1200 values for each of the three years in question, 1855, 1917 and 1965. The large grid squares can be looked upon as service regions or coherent regions from the point of view of commuting to work, their size being about 20—28 km from the periphery of the grid square to its centre. The combination of the original grid squares to make large grid squares helps to eliminate a number of the errors which are present in the original maps, errors which are the result of the population being distributed over the surface of a parish according to a rule of thumb. Such rules are that the population is concentrated in the vicinity of the parish church or that population distribution follows the distribution of dwelling houses.

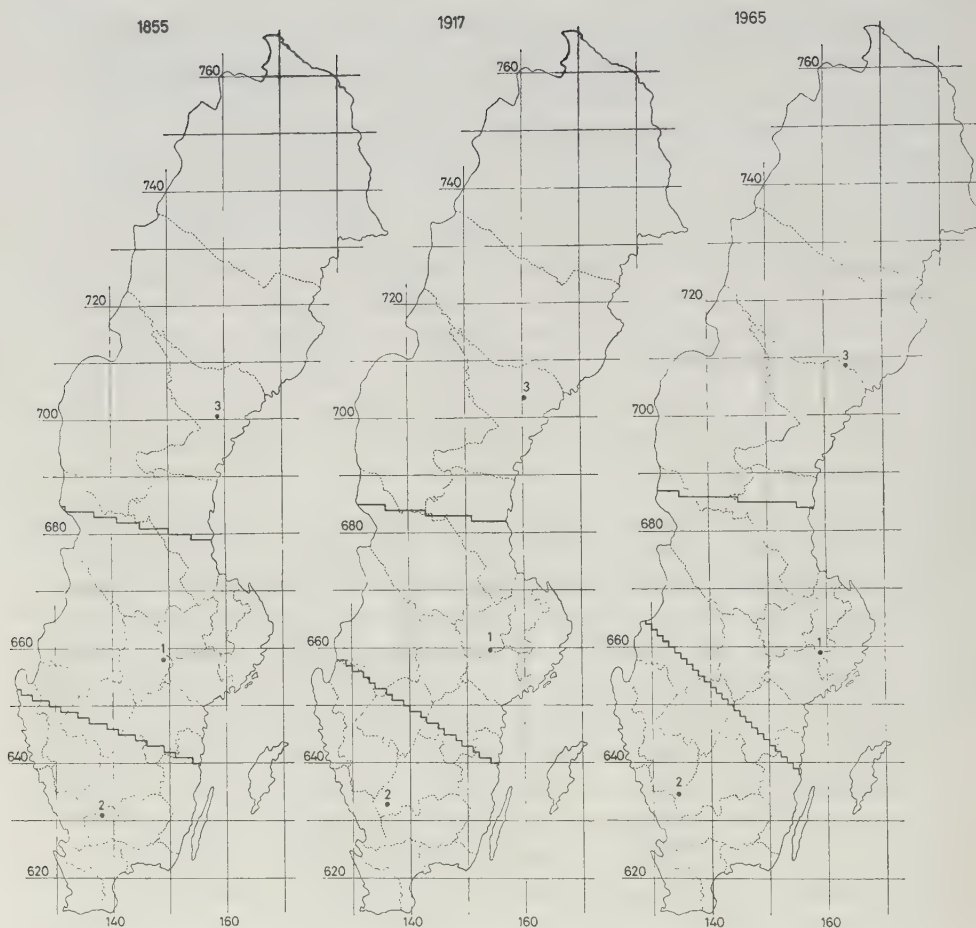


Figure 8.1. The optimum locations and umlands for three facilities. The arrangements have been obtained using the location algorithm, in conjunction with population data for the years 1855, 1917 and 1965.

8.2 The location of three facilities

The maps presented in chapter 4 make it possible to study how the alterations in the distribution of population which have occurred since 1855 effect the results from the location algorithm. NORLOC was used with the three maps as input data and with different numbers of facilities to be located. The first example involved the choice of locations for three facilities. Of the three regions thus obtained the northern region shows itself to be the most stable with respect to the boundaries of its umland. It can be seen from figure 8.1 that the southern boundary of the northern

F	Capacity in 1000's			Mean distance km		
	1855	1917	1965	1855	1917	1965
1	1551	2333	3431	123	111	96
2	1678	2496	3323	122	129	128
3	374	810	958	181	184	194
All	3603	5640	7712	129	129	122

Table 8.1. The capacity of each facility and mean travel distance incurred when three facilities are optimally located. The input data to NORLOC was in the form of square net maps showing the distribution of population in Sweden in 1855, 1917 and 1965.

region only moves about 30 to 40 km during the whole period from 1855 to 1965, the direction of the boundary remaining the same. Facility F_3 moves 70–80 km to the northeast during the period in question. During the same period facility F_1 moves 100 km to the east and facility F_2 50 km to the northwest. F_1 has thus moved nearer to Stockholm at the same time as F_2 moved nearer to Göteborg. As a result of these altered locations the boundary between the two regions changes direction, from westnorthwest—eastssoutheast in 1855 to northwest—southeast in 1965. The boundary alterations alter the shape of the respective umlands, even though the locations of the two facilities have not been greatly altered. A closer study of figure 8.1 leads to the conclusion that the alterations in the shape of the umlands are the result of F_1 moving towards Stockholm. The move towards Stockholm is a result of the very large increase in population which has taken place in Stockholm between 1855 and 1965. A clear picture of the effect the alterations in population distribution have on the transportation costs incurred by consumers in visiting their nearest facility can be obtained by calculating the mean transportation cost or mean distance for every facility. It can be seen from table 8.1 that the mean distance for the whole of Sweden has sunk from 129 km in 1855 and 1917 to 122 km in 1965. The urbanisation of Sweden has lead to a reduction in the mean distance travelled of about 7 km. From a total point of view an improvement in the accessibility of service has taken place. From the point of the view of the individual regions this is not true, since the improvement having taken place solely within the region around F_1 . A deterioration in the accessibility of service has taken place in the other two regions.

8.3 The location of 5 facilities

Seen from a theoretical point of view the location of three facilities should lead to unstable solutions. The function $f(X_{F1}, \dots, Y_{F3})$ can in the case of

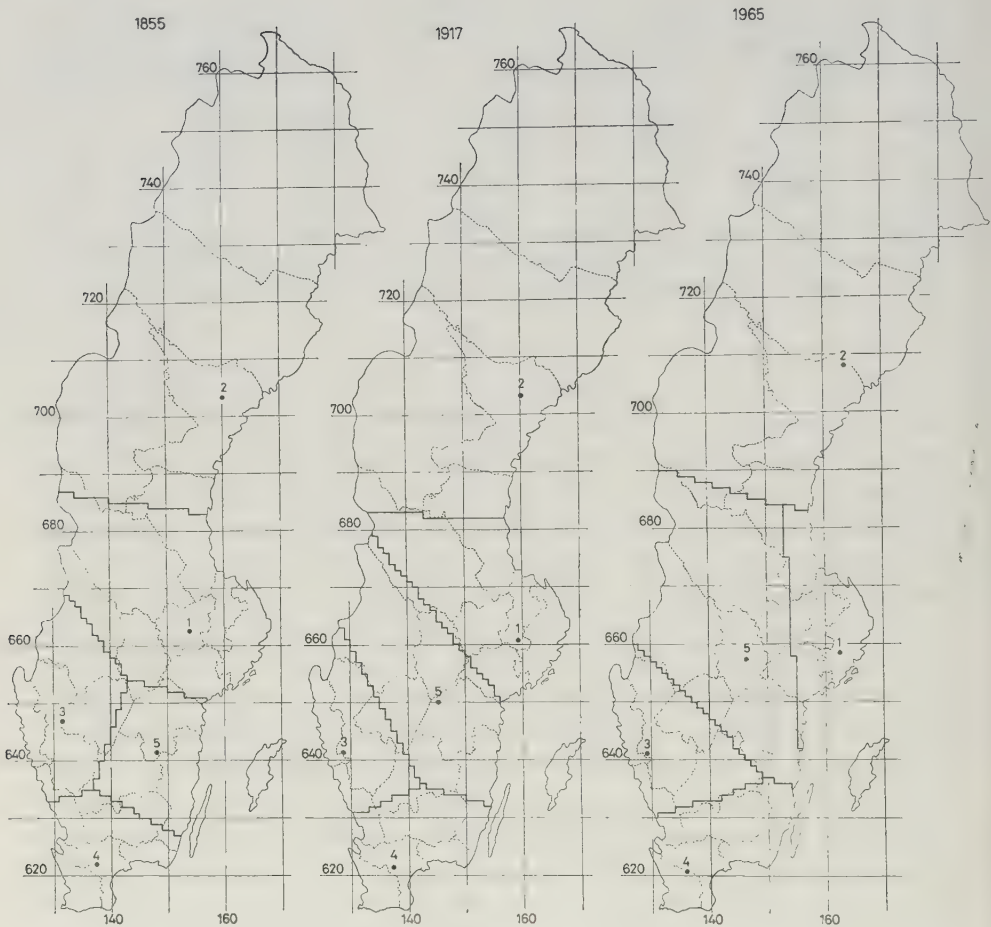


Figure 8.2. The locations and umlands for 5 optimally located facilities in Sweden in the years 1855, 1917 and 1965.

three facilities very easily lead to local minima. This due to the fact that the Swedish population can be divided into 4 groups from a chorological point of view. These groups are the three major city regions plus the north region. If 4 facilities had been located instead of three the locations chosen would have been very stable over time, compared with the situation with three facilities. An increase in the number of facilities located to 5 creates the situation where the number of facilities is 1 more than the number of natural regions. Thus the existing regions must be divided in a new way in order to give a new 5th region. The data from 1855 gives the 4 natural regions, northern Sweden (2), eastern Sweden (1), western Sweden (3) and southern Sweden (4), plus a new 5th region which can be referred to

F	Capacity in 1000's			Mean distance km		
	1855	1917	1965	1855	1917	1965
1	979	1486	1975	100	83	42
2	340	812	966	176	184	194
3	869	938	1586	82	59	63
4	808	1196	1497	71	78	79
5	607	1208	1689	78	99	102
All	3603	5640	7712	93	96	86

Table 8.2. The capacity of each facility and mean travel distance incurred when 5 facilities are optimally located.

as the central Swedish region. See figure 8.2. This new region consists of major parts of the counties of Kalmar, Östergötland and Jönköping. The names and the locations of the counties in Sweden are given in figure 7.16 on page 246.

The most stable of the 5 regions between 1855 and 1965 was northern Sweden. The centre of this region shifted 50 km to the northeast, at the same time as the centre of eastern Sweden moved to the southeast giving, as a net result, an unimportant alteration in the location of the boundary between the two regions. During the same period the centre of the western Swedish region moved nearer to Göteborg while the centre of the southern region remained relatively stable. The centre of the new central Swedish region moved considerably during the period 1855—1965, with the result that the boundaries of its umland also varied. Hence, in 1917 this region even included large parts of Skaraborgs, Örebro and Värmland counties together with the whole of the county of Östergötland and the northern parts of Jönköping and Kalmar counties. In 1965 the centre of the new region was near Örebro. See figures 7.16 and 8.2. This town can be assumed to be the natural centre of a central Swedish region consisting of the whole of Östergötland, Värmland, Örebro and Kopparberg counties, together with parts of Kalmar, Skaraborg, Sörmland and Västmanland counties.

Table 8.2 shows the capacity and mean transportation distance when 5 facilities are located for each of the three dates 1855, 1917 and 1965. An inspection of this table shows that the mean distance from *F*1 to all its customers is reduced considerably between 1855 and 1965, from 100 km to 42 km. In the case of the region around facility *F*3 the mean distance is reduced by 23 km from 1855 to 1917, but it is increased by 4 km between 1917 and 1965. In the case of all the other regions the mean distance increases between 1855 and 1965. This effect is particularly noticeable in region number 2, a result of the increased exploitation of the iron ore

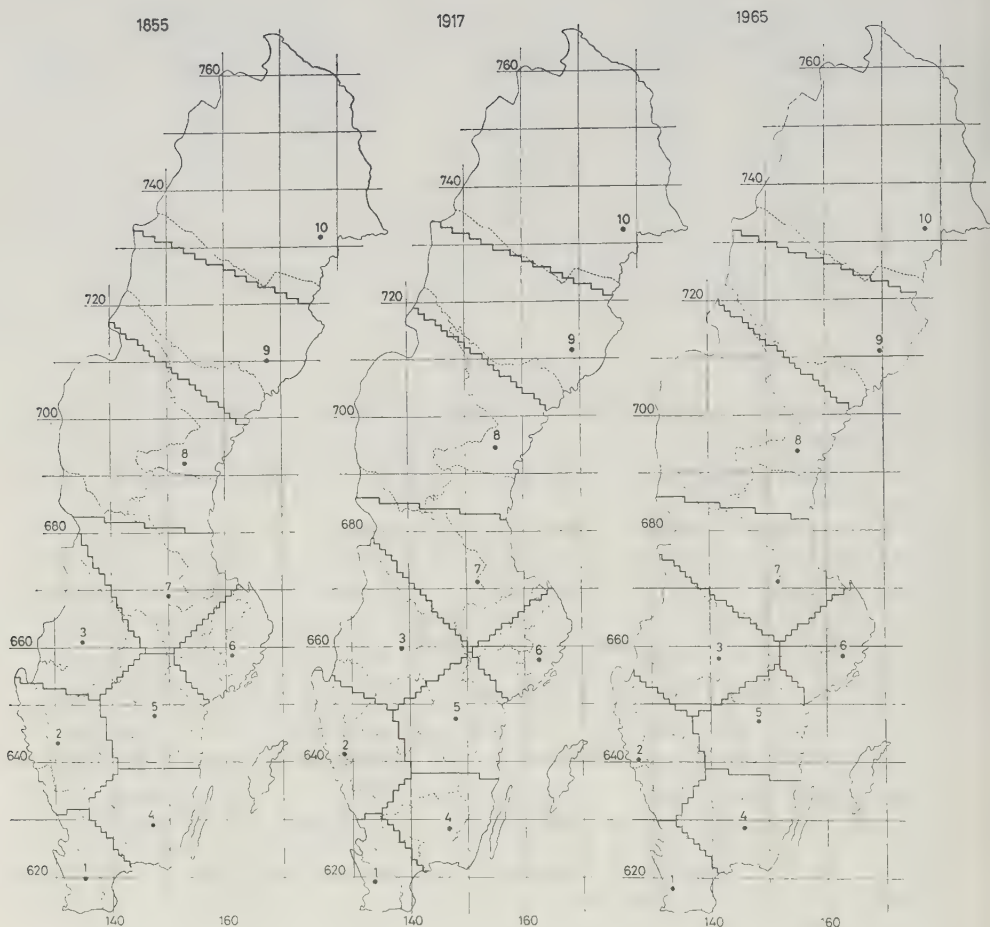


Figure 8.3. The locations and umlands for 10 optimally located facilities in Sweden in the years 1855, 1917 and 1965.

reserves in the region and the growth of residential settlements which has been associated with it.

One can summarize the above by saying that the use of NORLOC to locate 5 facilities produces results which are unstable with respect to population changes over time. This is particularly true of the 5th facility, the other facilities being relatively stable, with the exception of the movement towards Stockholm and Göteborg of F_1 and F_3 respectively. The large amount of movement experienced by F_5 during the period 1855—1965 results in major alterations in the shape of its umland and therefore in the umlands of the other facilities.

F	Capacity in 1000's			Mean distance km		
	1855	1917	1965	1855	1917	1965
1	533	759	984	49	45	41
2	657	904	1261	62	55	44
3	294	497	690	58	69	66
4	482	538	599	63	64	66
5	522	677	748	63	64	62
6	422	917	1832	50	33	29
7	327	543	623	67	68	65
8	195	426	423	90	87	85
9	99	197	280	82	86	85
10	73	183	271	85	92	94
All	3603	5640	7712	62	59	53

Table 8.3. The capacity of each facility and mean travel distance incurred when 10 facilities are optimally located.

The conclusion can be drawn that the distribution of population in Sweden is such that a minor alteration in it causes major variations in the regions obtained when locating 5 facilities. The actual variations in the distribution of population over time are not the major cause behind the different locations chosen for the years 1855, 1917 and 1965. The major cause lies in the fact that 5 facilities are to be located, the same effect can be noticed when 3 facilities are to be located. Both of these effects are a result of the fact that Sweden can be divided into 4 natural regions, attempts to locate a greater or smaller number of facilities increase the chances that local minima will occur.

8.4 The location of 10 facilities

The choice of locations for 10 facilities using NORLOC produces very interesting results. The input data used was the same as in the previous examples, the population distribution in Sweden for the years 1855, 1917 and 1965. It can be seen from figure 8.3 that approximately the same locations and umlands are obtained at each of the three points in time. Facility F3 moved the greatest amount, about 70 km to the east. The movement of F3 does not cause any major alteration in the arrangement of the umlands, the arrangements produced are therefore relatively stable.

If Sweden had been divided into 10 counties in the year 1855 using the principle described here, i.e. that the total cost of transportation the individuals to their county town is minimized, the result would have been an arrangement of counties capable of functioning satisfactorily even in 1965. It can be seen from table 8.3 that only the three most northern

regions had a population of less than 600 000 persons in 1965. If one introduces the restraint that every county must have at least 600 000 inhabitants the three most northerly counties would need to be combined. The resulting large county would be approximately the same as region 2 in figure 8.2. The NORLOC program does not contain possibilities for specifying restraints as to the size of regions. The same effect can be obtained by combining or dividing regions in a further processing stage. The introduction of this type of restraint into the program, i.e. that each region must have a minimum or a maximum population, is of course possible, but at the cost of a more complicated program more prone to fasten at local minimum points.

8.5 The location of 15 facilities

A repetition of the above experiment with 15 facilities instead of 10 results in a series of unstable arrangements. The function $f(X_{F1}, \dots, Y_{F15})$ has at least three different local minimum points, the three different minimum values being practically the same. In the case of the year 1855 the difference between the best and the worst solution is only 0.2 %. Interference between the two solutions occurs too. The second best solution for the year 1855 was found to be the best solution for the year 1917. The changes in population distribution between 1855 and 1917 were sufficient to alter a local minimum to an absolute one. An absolute minimum is, of course, taken to be the best of the collection of minimum points discovered.

Figure 8.4 shows the absolute minimum locations for 1965 and 1917 and the corresponding local minima for 1855. It can be seen from the figure that F_6 moves the most during the period, 60 km to the southeast. F_2 , F_4 and F_{10} move 40—50 km during the same period. The overall impression one obtains is that the arrangements produced are relatively stable, with respect to those alterations in the distribution of population which have occurred between 1855 and 1965.

8.6 The location of 23 facilities

At the present moment the Swedish mainland is divided up into 23 counties. It is therefore of interest to be able to compare this with the arrangement of 23 regions produced by the NORLOC program. NORLOC's arrangement is shown in figure 8.5 and table 8.4. It can be seen from the figure that the arrangements produced are very stable throughout the whole period 1855—1965.

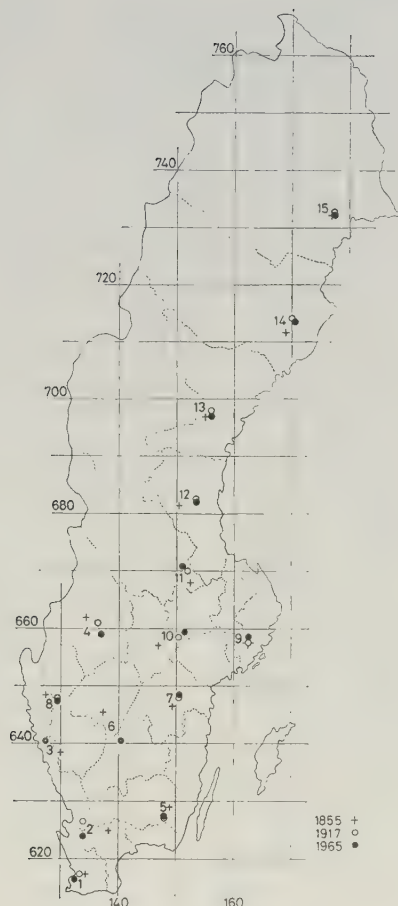


Figure 8.4. The optimal locations for 15 facilities in Sweden in the years 1855, 1917 and 1965.

From a population point of view the regions shown in figure 8.5 are somewhat more evenly balanced than the corresponding county arrangement. It can be seen from table 8.4 that region 21, which is the smallest region from the population point of view, has a population of 150 000, about 20 000 more than for the corresponding county. The largest region, around Stockholm, had a population of 1 327 000 in 1965, the corresponding population according to the present county arrangement being a little more than 1 378 000 persons. The range of size of the different counties measured in population terms is therefore greater than is the case for the 23 regions produced by NORLOC.

During the period 1855—1965 the mean distance between a customer's home and his nearest facility has been shortened from 38 km to 31 km. The figure for 1965 is too low, as it is based on parish data with the whole

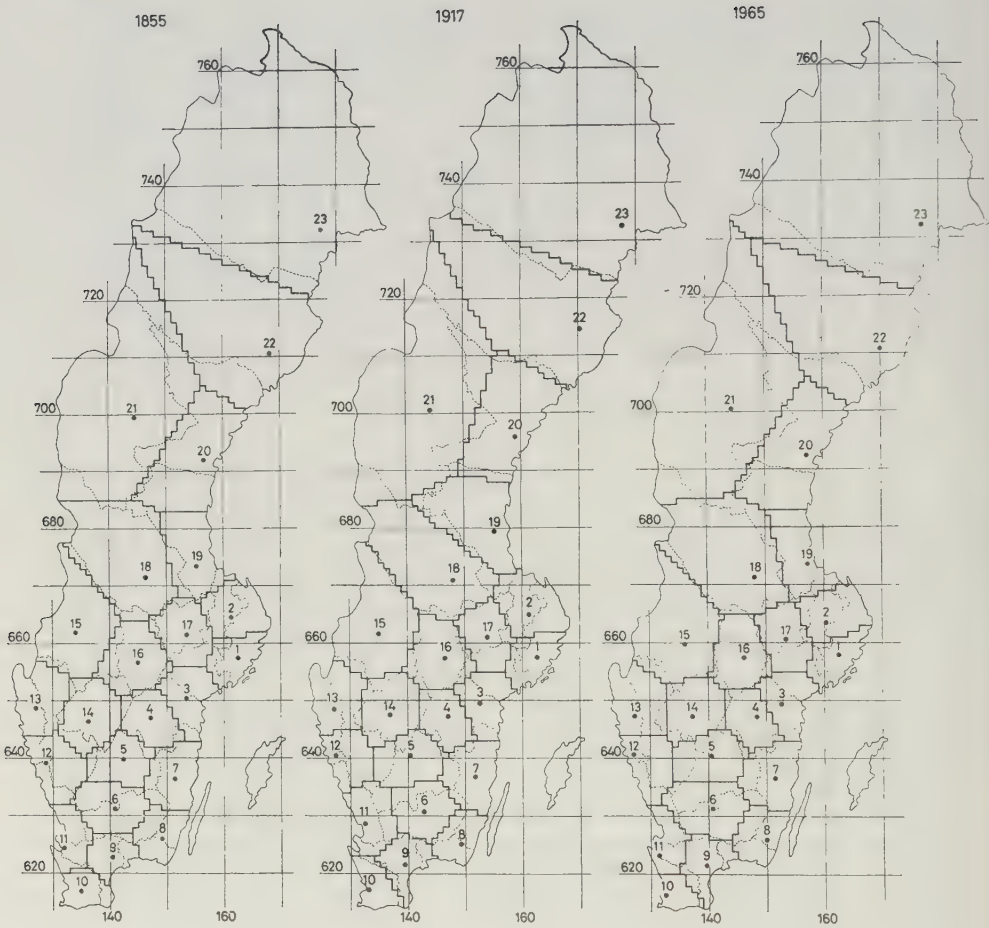


Figure 8.5. The optimal locations and umlands for 23 optimally located facilities obtained with the help of NORLOC. Input data: the distribution of population in Sweden in 1855, 1917 and 1965.

population located in the vicinity of the parish church. This is not the case in practice and mean travel distance is therefore longer.

The mean travel distance between customer and facility can be used as a measure of a region's service accessibility. The longer the mean transportation distance, the lower the level of service accessibility, and vice versa. It can be seen from table 8.4 that the service accessibility throughout the country has been improved by about 15 % between 1855 and 1965. The effect the increased level of urbanisation has had on the accessibility of service is restricted, seen from the transportation cost point of view. A further effect which should be noted is that mean distance in the Stockholm

F	Capacity in 1000's			Mean distance km		
	1855	1917	1965	1855	1917	1965
1	185	593	1327	15	8	11
2	131	189	237	34	37	26
3	153	220	260	34	27	27
4	153	168	223	28	27	24
5	121	193	229	33	31	26
6	128	146	204	37	33	39
7	146	165	165	40	42	41
8	192	206	208	35	35	38
9	157	221	262	27	32	31
10	261	463	517	28	27	14
11	157	213	326	33	37	33
12	253	435	812	36	18	21
13	216	240	264	37	37	31
14	218	228	234	34	33	29
15	211	242	303	46	48	45
16	175	252	284	36	33	27
17	147	243	368	35	31	27
18	160	261	295	51	49	45
19	95	228	223	45	53	35
20	118	253	278	58	51	49
21	66	141	152	63	69	64
22	92	170	269	77	77	80
23	70	171	272	84	90	94
All	3603	5640	7712	38	36	31

Table 8.4. The capacity of each facility and mean travel distance incurred when 23 facilities are optimally located, in Sweden, for the years 1855, 1917 and 1965.

region has increased from 8 km in 1917 to 11 km in 1965. The city of Stockholm has now become so large that any increase in population leads to an increase in the mean transportation distance, and a corresponding decrease in the accessibility of service.

A comparison between the present county boundaries and the NORLOC regions provides much of interest. It can be seen from figure 8.5 that the arrangement produced by the model corresponds very closely with the existing county boundaries. This is especially noticeable in northern and central Sweden, i.e. regions 14—23. Those government servants who divided northern Sweden into counties during the 1760's carried out their job in a very satisfactory manner.

The county division in southern and central Sweden was carried out during the 17th century, with minor alterations in the 18th. The county division is based on a considerably older "landskap" division. A number of provinces from the previous division were given the status of counties, amongst these were Östergötland, Värmland and Dalarna. See figure 7.16. Also three provinces, regions captured from Denmark and Norway (Blekinge,

Halland and Bohuslän) remained as counties in the new division. Småland was divided up into three counties. Dalsland was combined with the southwestern part of Västergötland to form Älvsborg county.

Several counties in southern Sweden are narrow and lie parallel with the coast. This is a result of the fact that coastal shipping was a widely used transportation method, it being easier to travel along the coast than direct inland. The choice of a long narrow form for these counties was sensible when the division was carried out. The transportation costs involved in visiting the county town were minimized as a result. At the present date the choice of this form is much less sensible, due to the fact that coastal passenger traffic is relatively much less important. The differences between the arrangement produced by NORLOC and the existing county boundaries were therefore only to be expected.

One can conclude by saying that the division of Sweden into 23 counties was carried out in such a way as to give a pattern acceptable even today. The differences between the region developed by NORLOC and the counties can be mainly explained by the changes in transportation which have taken place since the county reform was carried out. It is, of course, possible to believe that the original boundaries were not at first rational, but that over the years the central areas of the counties have expanded more than the areas nearer the boundary. If such was the case the division into regions given by NORLOC would not be as stable with respect to time as it actually is. The changes in population distribution between 1855 and 1965 have had little effect on the locations chosen. Changes in the distribution of population occurring before 1855 were very small. This can be checked, using the population records of the Swedish parishes, which go back to the middle of the 18th century. The population of the whole country increased before 1855, but the distribution of population increase was relatively even. The distribution of population in 1855 was much the same as it had been 100 years earlier. If data had been prepared for 1750 it would have given a similar arrangements to that obtained for 1855. The division of the country into counties would therefore appear to have been sensible when first carried out. The government servants responsible must have had good knowledge of the distribution of population and other relevant factors when the division was carried out.

8.7 Conclusions

The location model has shown itself to be relatively insensitive to the altered distribution of the Swedish population between regions which has

come about since 1855. In those cases where unstable solutions occur they are a result of the fact that function $f(X_{F1}, \dots, Y_{FN})$ has a number of minimum points and that interference occurs between them. In cases where unstable solutions occur the difference between the solutions is fairly small. This feature of the model makes great care necessary when applying the model to actual problems, due to the risk that solutions which are as good or nearly as good as that the model gives are disregarded.

A classification of changes in population distribution into changes occurring between regions and changes occurring within regions, depends to a certain extent on the size of region used. If a county is assumed to consist of a number of local authority areas then changes can occur between the local authority areas without causing changes in the population of the region relative to other regions. If one looks at an even lower level in the hierarchy, at the town or built-up area, it is possible to say that the construction of new housing estates causes the distribution between regions to be altered. In such cases the location model does not give stable solutions when data from different points in time are used as input.

9 Stepwise regionalization

The choice of locations for a number of facilities involves two stages, the first being the choice of a location for each facility, the second the division of the area into regions. Cooper (1963) used this division in his location algorithm. He alternates between the definition of each region and the choice of a location for the facility in the region. The location problem can be defined as the division of an area into a number of optimal regions and an optimum location for facility then being chosen within each region (Scott 1971). This approach to the statement of the problem reduces the multiple location problem to the simpler problem of choosing the position of one facility. This problem was handled by Weber in 1909 and is therefore called the Weberian location problem (Weber 1958). The division of an area into a number of regions can be done in a number of ways including nearest neighbour, cluster analysis and tree searching methods (Bunge 1966, Scott 1971). Against this approach can be used the fact that a multiple location algorithm can also be used as for the division of an area into regions.

Practical planning problems are often of a type which require the hierarchical division of an area into main regions and sub-regions. This can be done in two ways, from the top of the hierarchy downwards or from the bottom of the hierarchy upwards. In the first case, called here stepwise regionalization, the main regions are defined first and then divided into sub-regions. In the other case, called here combinational regionalization, the sub-regions are defined first and then a certain number of them are combined to form each main region.

The larger the number of facilities to be located, the more complicated the total transportation cost function and the more the risk that the function has more than one minimum point. A large number of facilities to be located can force one to use the first approach, where the main regions are defined first. A comparison of the results obtained using the two approaches to the same problem was also considered to be of interest.

The map shown in figure 6.10 illustrates the division of Sweden into 68 regions. The population of each region was then assumed to be located at

	Stepwise regionalization	Combinatorial regionalization
7 main regions	100	101.7
68 sub-regions	101.4	100

Table 9.1. The relationships between the total transportation costs when Sweden is divided into 7 main regions and 68 sub-regions using the two different regionalization models; the stepwise model and the combinatorial model.

the centre of each region. This resulted in a square net map of Sweden with a maximum of 68 populated squares. This map with a grid size of 100 km² was used as input data to NORLOC, the result being a division of Sweden into a number of major regions, in this case 7. The result from this execution of the NORLOC program is shown in figure 9.1 A. The division of Sweden into 7 regions was repeated using NORLOC with the maps shown in figures 4.5 and 4.6 as input data. The result from this execution of the program is shown in figure 9.1 B. A comparison of the two maps shown in figure 9.1 allows one to say that the locations of the 7 facilities and the division of the country into regions are for practical purposes identical. This means that results obtained by the combination of smaller sub-regions to make main regions and the division of the area into main regions directly are for practical purposes the same. In the case described here the stepwise method produces a result about 1.7 % more efficient than the combinatorial method. See table 9.1.

Each of the 7 main regions shown in figure 9.1 B is also to be divided into a number of sub-regions. The number of sub-regions in each main region was determined by placing the map with 7 regions over the map with 68 regions and by counting the number of sub-regions falling within each main region. The NORLOC program was then used to produce a new division of the country into 68 sub-regions, the boundaries of each group of sub-regions coinciding with the boundaries of the main regions. This was done using a series of square net maps, each one describing a main region, which were prepared using the computer at the same time as the division into main regions was carried out. Figure 9.2 B shows the results from these 7 executions of the NORLOC program combined to form one map. This division of the country into 68 regions is not optimal. Certain of the sub-regions in figure 9.2 B differ considerably from the corresponding sub-region in figure 9.2 A. This is especially noticeable in the case of sub-regions 14, 40 and 7. As is seen from table 9.1 the total transportation cost, measured at the sub-regional level, is only 1.4 % higher using the stepwise approach than that obtained when using the direct approach.

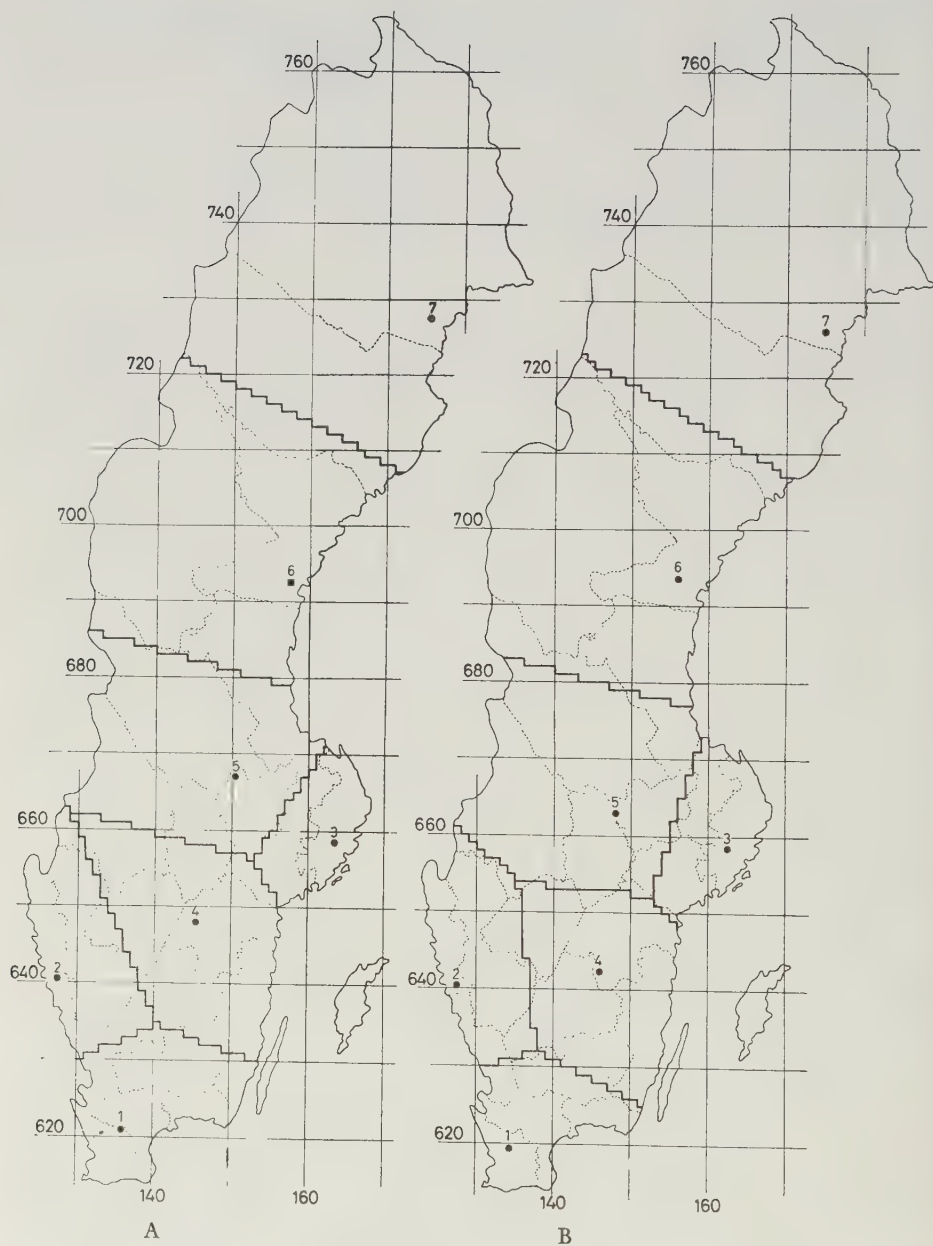


Figure 9.1. A division of Sweden into 7 regions, performed using NORLOC. Fig. A shows the regions obtained when 68 sub-regions are combined to form 7 regions, fig. B shows the result obtained when the country is divided into 7 regions in one operation.



Figure 9.2. A division of Sweden into 68 regions. Fig. A shows the result obtained when the division was carried out in one operation, fig. B shows the result obtained when each of the 7 regions in fig. 9.1 B is divided up into a suitable number of subregions.

The larger the number of regions an area is to be divided into the longer the time the computer requires to carry out the division. In the previous example the computer time required to divide Sweden into 68 regions was about 180 minutes. The alternative approach, where the country was divided into 7 regions, each region then being further divided into a number of sub-regions, required only 6 minutes. Both of these times refer to the Univac 1108 computer at the Lund Computer Centre. As the result from both of these approaches is practically identical the faster stepwise approach using 2 stages would be used for practical problems. A greater number of stages can also be used, a division into main regions, regions and sub-regions is possible to imagine. The effectiveness of the stepwise approach makes possible the handling of large problems with short execution times. A further advantage of this approach is that it is possible to control the size of the regions at the next lower level in the hierarchy. That means that one will always obtain sufficiently large regions at all levels.

10 Stepwise localization

In practical planning situations, it is often the case that the locations of a number of facilities are already decided. The problem is then to locate a number of new facilities in optimum locations, with respect to population distribution and the location of existing facilities. This situation occurs often so that a location model ought to be designed so as to be capable of holding certain facilities fixed while others are allowed to move. NORLOC has this capability; certain facilities can be held fixed and the remainder allowed to move.

10.1 A comparison between stepwise and optimum localization

In many cases the number of facilities which are to be built is unknown. The problem is therefore commenced by the choice of a location for only one facility bearing in mind the distribution of population within the area. After a period of time the discovery is made that one facility is insufficient and a decision is made to build one more. The location of this second facility is chosen with respect to the distribution of population within the area and the location of the first facility. After a further period has elapsed the discovery is made that two facilities are insufficient and a decision is made to build a third. The location of this one is chosen with respect to the distribution of population and the locations of the other two facilities. It is therefore possible to say that the choice of locations is carried out in a number of steps. In the case described here only one facility was located at a time. Using NORLOC it is possible to locate different sizes of groups of facilities at every step.

The approach to the choice of locations which is used by practical planners is often of the type described above, i.e. a technique with the locations being chosen in steps. The results obtained, using such an approach, can be highly unsatisfactory, even to the point where the locations chosen prohibit a later increase in the number of facilities. One can say that the available

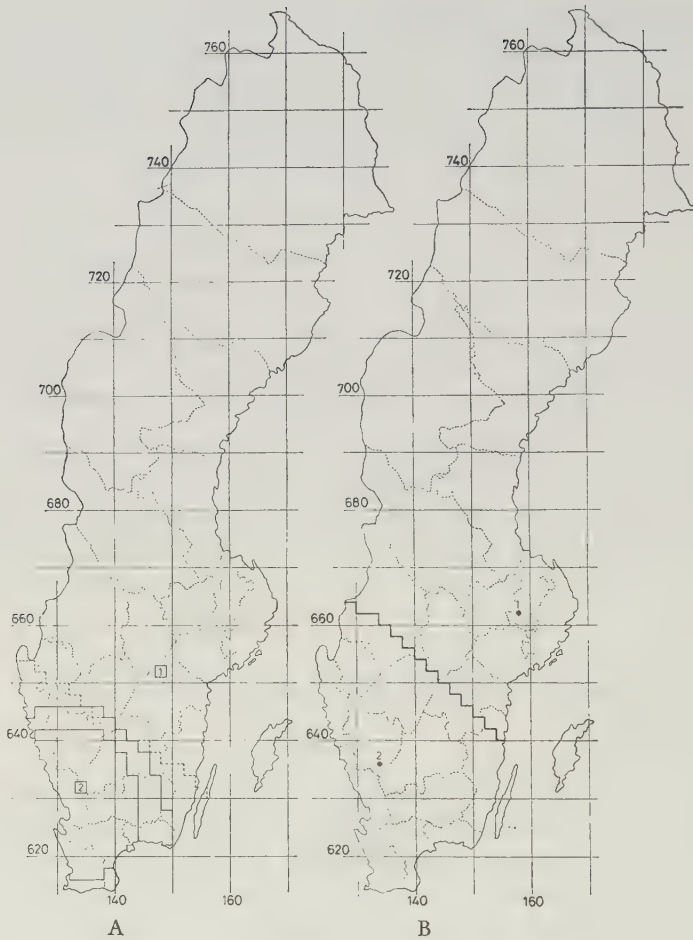


Figure 10.1. The locations and umlands for two facilities, fig. A shows the results obtained using one fixed and one moveable facility, fig. B shows the optimum situation with both facilities moveable.

resources are not used in an effective manner. The following example illustrates how strange the result can be when locations are chosen in stages.

10.1.1 Locating two facilities

The first stage involves the choice of location for one facility, which is to serve the whole of Sweden. This facility locates itself to the northwest of Finspång in a sparsely populated area. See figure 10.1. Table 10.1 shows that the mean travel distance between all the customers and the facility is 248 km.

I	Stepwise	Optimum	Best of 10- <i>I</i> units
1	248	248	255
2	194	165	195
3	141	123	154
4	110	100	118
5	98	87	93
6	85	76	78
7	76	68	70
8	68	61	63
9	62	57	58
10	58	54	54

Table 10.1. The mean travel distance occurring when 10 facilities are successively located using stage by stage choice of location and the "best of 10-*I*" method.

If one is prepared to allow this distance to increase by up to 5 % the facility can be located in either Linköping, Norrköping or Örebro. An increase in the mean distance of 10 % allows the facility to be located in either Stockholm, Jönköping, Skövde or Västerås. The right hand map in figure 7.16 on page 246 shows the positions of some Swedish towns, including the towns mentioned above.

The next stage in the problem involves the choice of a location for the second facility, with the first facility remaining in its original location. The location chosen by NORLOC for the second facility is in southwestern Sweden, near the boundaries of Halland, Småland and Västergötland. This is also a sparsely populated area, the nearest town being Halmstad. The mean transportation distance then sinks to 194 km, see table 10.1. If one is prepared to allow this distance to increase by up to 5 % the facility can be located anywhere within a wide area of southwestern Sweden. It can be seen from figure 10.1 that the towns of Göteborg, Malmö and Växjö are possible alternative locations. All of these towns lie within the 5 % isarithm. This isarithm is shown in figure 10.1 A by the inner of the two polygons surrounding facility number 2. The outer polygon is equal to the 10 % isarithm.

If it had been known at the beginning, that two facilities were to be located then the technique using two steps would not have been chosen. Instead, both facilities would have been allowed to move and located in one and the same stage. It can be seen from figure 10.1 B that the location chosen for facility number 1 is completely different to that obtained using the stepwise approach. Facility number 1 now locates itself just to the northwest of Stockholm. The mean travel distance obtained, when both facilities are located in the same stage, is 165 km, compared with the 194 km obtained when the choice of location is done in two steps.

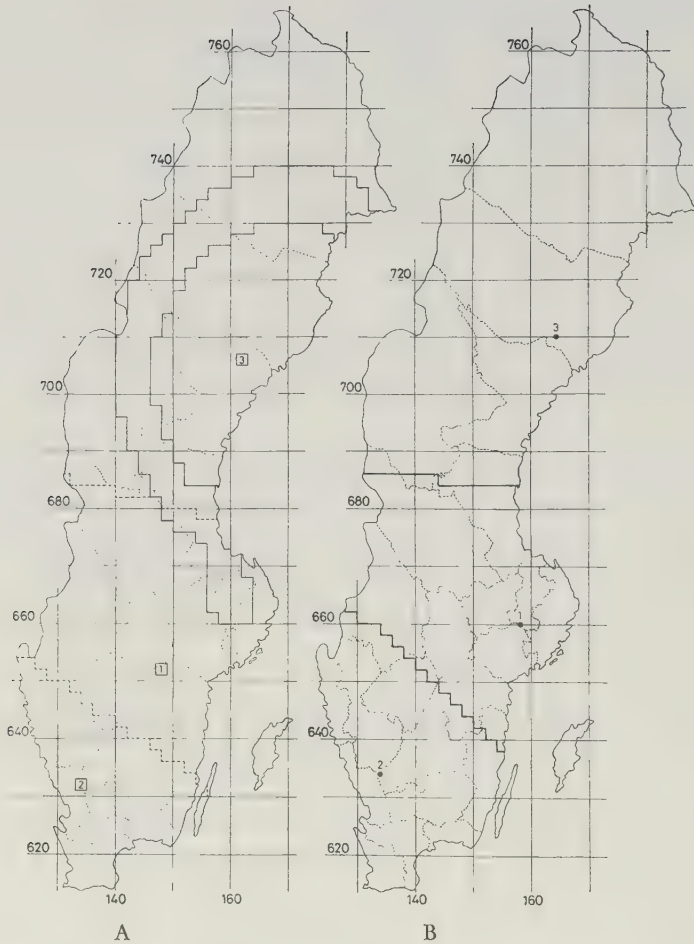


Figure 10.2. The locations and umlands for three facilities, obtained using stepwise choice of location (fig. A) and optimum choice of location (fig. B).

10.1.2 Locating three facilities

Continued application of the stepwise location technique gives a location for a third facility in the north of Sweden, just to the west of Umeå. The mean transportation distance is in this case 141 km. Figure 10.2 A shows the result, together with two areas within which the transportation distance is less than 5 % respectively 10 % longer than the minimum value (141 km). If a mean distance up to 10 % longer than the minimum is acceptable then the third facility can be located within a larger area, including Stockholm and the towns of Luleå and Östersund.

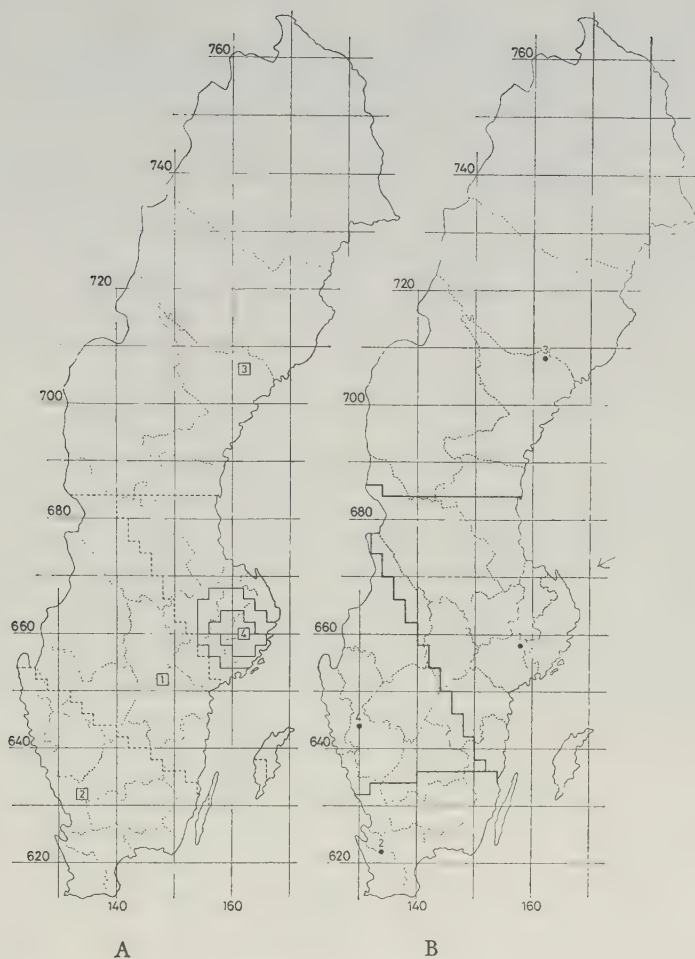


Figure 10.3. The locations and umlands for 4 facilities, obtained using stepwise choice of location (fig. A) and optimum choice of location (fig. B).

The optimum location of three facilities, with each facility able to move and location carried out in one stage, gives a 15 % reduction in the mean transportation distance compared with the stepwise location. A comparison of the two approaches shows that it is mainly facility *F1* which moves. This facility now lies in Stockholm (figure 10.2 B).

10.1.3 Locating 4 facilities

The stepwise location model gives a facility in Stockholm first with step number 4, see figure 10.3 A. By comparison the optimum location of

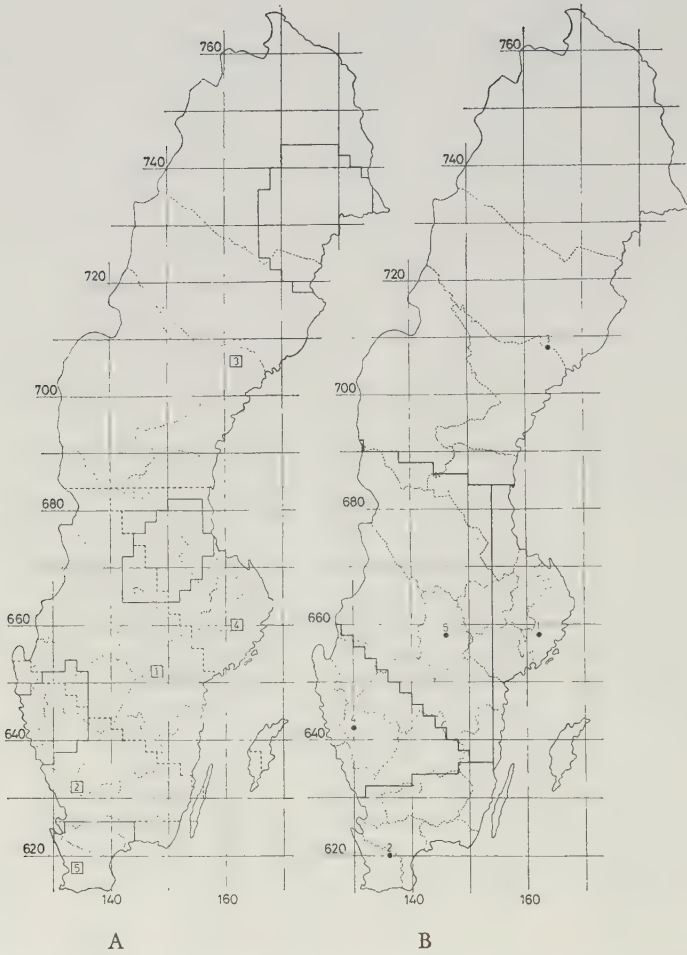


Figure 10.4. The locations and unlands for 5 facilities, obtained using stepwise choice of location (fig. A) and optimum choice of location (fig. B).

4 facilities retains *F1* and *F3* in Stockholm and Umeå. The southwestern region is divided into two regions, one lying near the south coast and the other near the west coast, see figure 10.3 B. Sweden can be divided into 4 natural regions using this approach, the north, Stockholm, Göteborg and Malmö regions.

10.1.4. Locating 5 facilities

It can be seen from figure 10.4 A that no less than 4 possible location exist for the 5th facility, if the choice of location is made in steps. A location in

Malmö gives the lowest transportation costs, but if a minor increase can be tolerated then location in Göteborg, Falun or Luleå are equally satisfactory. If, on the other hand, all the facilities can be moved, then 4 of them locate themselves in the vicinity of each of the three major towns plus Umeå. These are approximately the same locations as were obtained for facilities *F1*, *F2*, *F3* and *F4* in the case when 4 facilities were located using the direct method. The 5th facility locates itself in the vicinity of Örebro (figure 10.4 B). A comparison with figure 10.3 B shows that the Stockholm and Göteborg regions have become three regions.

10.1.5 Locating 6—10 facilities

Göteborg is chosen as the location for the 6th facility, when locations are chosen in steps, Falun is chosen for the 7th and Boden for the 8th. These locations are shown on the map at figure 10.5 as squares with the number of the respective facility in the centre. It can be seen from table 10.1 that for 2 and 3 facilities the transportation costs are about 15 % higher when location is carried out using the stepwise model. When 4 to 9 facilities are located this difference is reduced to about 12 %. The same effect can be noticed when the number of facilities is increased to 10. In this case, the difference between the two solutions is about 7 %.

The optimum locations for 7 facilities include a facility at Hedemora just to the south of Falun. In this case Falun would be the most logical location for the facility since Hedemora is too small a town for a facility of this kind. A further facility is located at Skänninge just to the west of Linköping. Hedemora and Skänninge are marked on the map at figure 10.5 by small circles containing an H and an S, respectively. See also figure 7.16. In the north of Sweden a facility is located just to the south of Boden, (*F8* in figure 10.5), another is located at Sundsvall. In addition, facilities are also located in the three major city regions. An increase in the number of facilities to 8 causes the facility at Skänninge to be removed and replaced by facilities at Växjö and Örebro. The facility at Hedemora is moved to Falun. The other facilities remain in their original positions. The locations chosen for 8 facilities using the optimum approach are marked by large circles in figure 10.5.

Figure 10.6 is a map showing the optimum locations of 10 facilities. A comparison between figure 10.5 and 10.6 shows that Umeå is chosen as the location for unit 10, while Örebro is replaced by Linköping (1) and Karlskoga (8). Karlskoga is a small town and it would therefore appear logical to move facility number 8 a little to the west to Karlstad.

In conclusion one can say that the stepwise choice of locations leads to much less than optimum result. It can be seen from table 10.1 that it is possible to provide the same accessibility of service using a smaller number of facilities in optimum locations as is provided by a greater number of facilities, the locations of which have been chosen using the stepwise model. A further feature of the experiments described above is that, in the cases 7, 8 and 10 facilities, the optimum location model gives 6 facilities the same locations. It is therefore possible to commence a building program and build this first 6 facilities without a definite decision being made about the final number of facilities to be built or their eventual locations.

10.2 The best of $N-1$ facilities

In many cases the time required to plan and construct a facility is much less than the time it will actually be in use. Further, it has been shown in chapter 8 that the alterations occurring in the distribution of population during the planning and construction period are unimportant from the choice of location point of view. At least this is the case in Sweden when locating facilities at a national or regional level. It is possible to look at location problems as if all the facilities were to be built at the same time, even if the construction program actually requires a number of decades. Such an attitude makes possible the use of the optimum approach to choice of location instead of the approach with a number of stages. A prerequisite for the use of this approach is that the number of facilities N is known in advance. The optimum locations can then be determined using NORLOC, as well as the order in which they should be built. Even if the period required for the building program is relatively short the facility to be built first should be that facility with the best overall location. This best overall location can be determined fairly easily, as the number of possible locations (N) is usually small. Table 10.1 illustrates a case when N is equal to 10. The first facility is, in this case, allocated to Linköping. The choice of Linköping as the location for the first facility gives a mean transportation distance which is only 3 % higher than that obtained when the facility is located in the optimum location for one facility.

In the next stage of the problem, when a location is to be decided for a second facility from amongst $N-1$ possibilities, there is already one fixed facility. One chooses a location from amongst the $N-1$ possibilities which minimizes the mean transportation distance. In the case of 10 facilities the second facility is placed in Sundsvall which lies in the centre of the northern Swedish region, see figures 10.6 and 7.16. The mean distance is

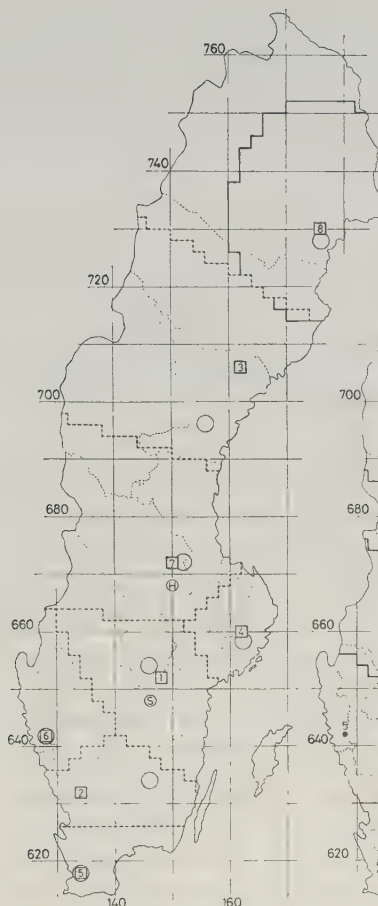


Fig. 10.5

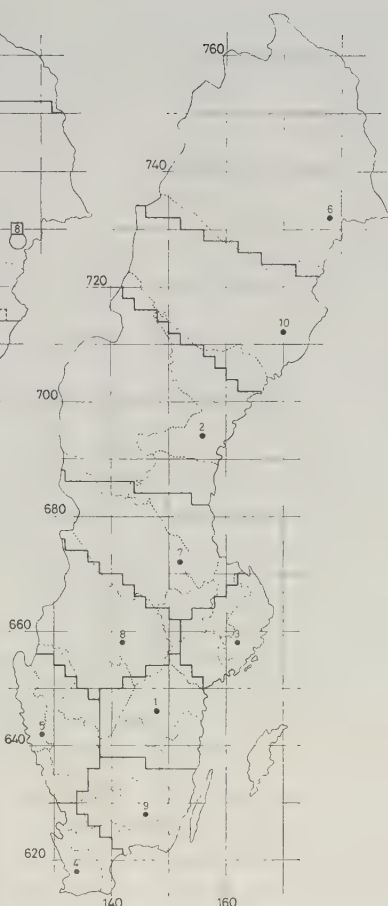


Fig. 10.6

Figure 10.5. The squares indicate the locations for 8 facilities, generated using stepwise choice of location. The boundaries of the umlands are shown by dotted lines. The circles indicate the locations for 8 facilities chosen using the optimum method.

Figure 10.6. The order of location chosen for the optimally located facilities using the "best of 10-1" method.

reduced to 195 km, which is very near the distance obtained when locating two facilities using the stepwise location model.

The procedure is repeated for the remaining $N-2$ facilities. It can be seen from table 10.1 that after step number 4 the mean distance becomes less than that obtained using the stepwise approach. This improvement is of such a size as to bring the mean distance into line with that obtained when locating the same number of facilities using an optimum approach.

Facility <i>I</i> +1	town	Capacity
1	Linköping	10
2	Sundsvall	5
3	Stockholm	24
4	Malmö	13
5	Göteborg	16
6	Luleå	3
7	Falun	8
8	Karlskoga	9
9	Växjö	8
10	Umeå	4

Table 10.2. The order of each facility when locations are determined using the "best of 10-*I*" method. The capacity is given as a percentage of the Swedish population in 1965.

As can be seen from figure 10.6 and table 10.2 the use of stepwise localization causes the three major cities, Stockholm, Malmö and Göteborg to receive facilities number 3, 4 and 5 respectively. The 6th facility is located at Luleå in the north of Sweden. When the choice of locations was carried out stepwise Umeå was allocated facility number 3. The "best of 10-*I*" model gives facility number 10 to this town. This is an interesting result, as Umeå was chosen as the location for the first university to be built outside of the three major urban regions. Further university facilities have been built at Linköping and Växjö. Karlskoga is too small a town to be chosen as the location for a university. As a result the Karlskoga region has been divided up between Karlstad and Örebro with a branch university in both of these towns. A new technical university is being built in Luleå. All of the 10 locations discussed in this example have been allocated a university or branch university with the exception of Sundsvall and Falun.

Table 10.2 shows that the Stockholm region contains the largest proportion of Sweden's population, 24 %, this being about 2 000 000 persons. Göteborg and Malmö taken together have 29 % of the population, about 2 400 000 persons. Taken together the three major urban regions contain more than half of the Swedish population. The other regions in southern and central Sweden, which were obtained using NORLOC to locate 10 facilities, have populations between 650 000 and 800 000, these being the Växjö, Karlskoga, Falun and Linköping regions. The other regions, which lie in the north of Sweden, have populations between 250 000 and 400 000 persons.

When the approach to choice of location used is based on the best of 10-*I* then the first location to be chosen for a facility is that in Linköping. The first facility should not have too large a capacity. Table 10.2 gives a capacity of 800 000 persons. When demand for the facility rises above

800 000 persons then the next facility, in Sundsvall, must be built. The size of each new facility is decided in relation to the situation which comes about when all of the 10 facilities have been built.

10.3 Conclusions

The stepwise approach to choice of location, where only the I th facility is free to move has obvious shortcomings and should therefore only be used in extreme situations. A far better approach to choice of location is provided when all the facilities are allowed to move, the locations for N activities being chosen in one operation. A prerequisite for the use of this approach being that N is known. In situations when the time required to plan and build the N facilities is short, compared with the useful life of the facilities, the order in which they are built is unimportant. On the other hand situations where the time required is of importance do occur. In such cases the location of each new facility is chosen from the "best of $N-I$ " locations. I is, in this case, taken to refer to the number of facilities, the location of which has already been decided upon. Choice of location according to the "best of $N-I$ " method gives a final result identical to that obtained when using the optimum location model.

11 Simultaneous hierarchical localization

Up to now the location model used has only taken into account the cost incurred by the user when visiting the facility. This type of model chooses locations which are good from a consumer's point of view. From the point of view of a business economist these locations chosen might be unsatisfactory. This due to the fact the business economist attempt to maximize the profits of his company or minimize its costs. The costs incurred by the user when visiting the facility do not enter into his calculation, as long as the number of customers visiting the facility is not reduced to a too great degree. Using such calculation methods one can show that large, out-of-town shopping centres and supermarkets are the cheapest and most effective way of distributing goods. When the total cost to the community of such an approach to the distribution of goods is being calculated the consumers' transportation costs must be included, as well as the costs incurred by the community in building and maintaining the extra road facilities which become necessary. Further factors which enter into the calculation are the costs incurred by third parties. An example of such a cost being the increased time a person needs to spend in travelling a route which passes in the vicinity of an out-of-town shopping centre, such shopping centres being heavy generators of traffic. A calculation of the type described is, of course, exceedingly difficult to carry out. Some simplifications must therefore be made. It would seem reasonable to assume that every individual should, as a right, have good access to a certain range of facilities. The simplifications one makes should, therefore, not exclude the costs incurred by the user when visiting the facility. When a decision is to be made between a location which is good from the business economist's point of view and a location which is good from the user's point of view there is no prior rule which states that the first alternative should be chosen at the expense of the second one.

11.1 Costs incurred in the transportation of raw materials

Törnqvist (1963) investigated industrial firms' transportation costs and classified them into two types, the costs incurred in transporting raw materials to the factory and costs incurred in transporting finished products from the factory to the customer. Such an investigation requires two types of information, a description of the spatial distribution of the customers and description of the spatial distribution of the sources of raw materials. The latter can be given in the square net map form or in a coordinate table. A coordinate table would appear to be the best alternative as the number of sources of raw materials is usually small in number. The introduction of two sorts of transportation costs into the model makes a decision regarding the relationship between them necessary. In principle, the introduction of transportation cost does not seriously complicate the model. Further transportation cost functions with their corresponding maps or tables can be introduced into the model if necessary. The only drawback with such an approach is that the risk is increased that the function $f(X_{F1}, \dots, Y_{FN})$ has more than one minimum point.

11.2 The simultaneous location of two types of facility using a single map of population distribution as input data

Chapter 9 discussed two different ways of dividing Sweden into hierarchical regions. One approach divided the country into 68 sub-regions which were then combined to form 7 main regions. The alternative approach carried out the division into 7 regions first. The main regions were then divided into a suitable number of sub-regions. Both approaches gave solutions where the centres of the main regions at the same time were centres of sub-regions. A complicating factor when using these two approaches is provided by the fact that it is impossible to determine which of the two alternative arrangements is the best. This is due to the fact that one alternative is best when the calculation is carried out at the sub-regional level, the other being best when the calculation is carried out at the level of the main regions. The only way to obtain an optimum solution to the problem of locating both sub-regional centres and the centres of main regions is to make the choice of locations simultaneously. The problem becomes one of deciding the location of 68 sub-regional centres, 7 of which also function as the centres of main regions. The individual

customer is assumed to make visits to centres of both types, maybe the visits to the main regional centre are less frequent than visits to the sub-regional centre.

A drawback incurred when extending the location model, in the way described above, is the increased risk that the choice of locations fastens at a local minimum point. A further drawback is that the computer time required when locating different types of facility simultaneously is very long. This approach to the choice of locations is the most expensive of all the alternatives described. A high price is paid for the increased accuracy in the choice of location which it provides.

11.3 The simultaneous location of a number of types of facility using a number of maps as input data

The hierarchical location model can be made increasingly complicated without any changes in its basic principles. The number of levels in the hierarchy of facilities can be increased, a special square net map of a spatial distribution being required for each new level. The model can be given a general character and be used for problems other than the hierarchical choice of locations. This is done, for instance, when the facilities are assumed to consist of different combinations of a number of service activities, the number of different service activities being referred to as HN . The problem is then to decide upon locations for a certain number of facilities containing different combinations of the available service activities. The number of facilities being referred to as N . The computer must be given information regarding which service activities are included in which facility, this being done using a matrix with N rows and HN columns. The computer must also be told which transportation cost function is to be used in calculating the components of the total transportation cost. Finally, the computer requires HN square net maps as input data. If the number of demand or supply points shown on a certain map is small it can be replaced by a table. This general version of the location model includes Törnqvist's model as a special case. This special case used a transportation cost function where both the costs of transporting raw materials to an industrial facility and the costs of the distribution of finished products to the customers were represented. Another special case of this simultaneous location model is that described in section 11.2, where only one map was used as input data. In such cases the map is given as input data HN times. This version of the model can also be used in the choice of locations for facilities which require a very long period for their completion. The different maps representing different points in time.

The need to allocate locations simultaneously to facilities at different levels in a hierarchy occurs very often in connection with practical planning. The choice of locations for schools provides a practical example of such a problem. The Swedish primary schools are divided into three categories, one for children between 7 and 9, one for children between 10 and 12 and one for children between 13 and 15. In order to avoid having to make small children travel long distances to school the first type of school is usually small, serving a restricted local area. The slightly older children can be allowed to travel a slightly longer distance to school, which results in the schools for children between 10 and 12 being larger than the schools for children between 7 and 9. Schools for children aged between 10 and 12 are not usually built in isolation. They are nearly always combined with a school for children aged between 7 and 9. At the next higher level in the hierarchy one finds the schools for children aged between 13 and 15. These schools are usually large with up to 150 pupils in each age group. A school of this type is usually combined with schools for the two younger age groups.

The choice of locations for schools can be described as consisting of the choice of location for a large number of schools for the youngest age group, some of these schools being independently located the remainder of the schools being combined with a school for the next age group, 10 to 12. Some of these schools form independent units but some of them are a part of a very large school with facilities for all three age groups. It can therefore be claimed that the choice of locations for schools provides a very good practical example of the simultaneous choice of location for facilities at different levels in a hierarchy.

This version of the location model is the most complicated. The risk that the total transportation cost function has more than one minimum point is extremely large. A further complication is that the computer requires an exceedingly long time to determine a solution to this type of problem. It is therefore vitally necessary to give the start locations of the various facilities as accurately as possible. The start locations must be chosen to lie as near as possible to the optimum locations.

Before a final decision is made regarding the use of the location model in this complicated version, an investigation should be carried out to see whether a simpler version could be applied. The school location problem could be simplified to the choice of locations for a number of schools for children in the oldest age group. This means that the local authority area is divided up into a number of high school districts. The next lower level in the hierarchy can then be tackled by choosing the locations for a number of schools for children in the middle age group within each high school

district. These locations are chosen with one of the schools located at the same place as the school for the oldest age group, the rest of the schools are free to locate themselves anywhere within the district served by the school for the oldest age group. The procedure is then repeated using schools for the lowest age group. This approach requires a fraction of the computer time necessary when choosing locations for schools at different levels in the hierarchy simultaneously. The solution one obtains using the economical approach is, of course, not optimum but very nearly so. Such a near optimum solution is, of course, completely acceptable. This due to the fact that even a theoretically optimum solution must be amended to bring it into line with the actual options open to the planner. In conclusion it can be said that whatever the model used, simple or complicated, it can never replace the planner. A model must be looked upon as an aid, capable of defining a range of alternative locations. The locations it gives can always be questioned from one point of view or another and the results it gives should therefore not be implemented blindly. Every model is a simplified representation of reality. No model, however complicated, can be considered so perfect as to make unnecessary a careful manual inspection of the results it gives and their adjustment with respect to the practical alternatives open to the planner.

12 The location model and planning at the regional and local levels

In the introduction to this report it was described how planning in Sweden is carried out at different levels, such as local, regional and national. The examples given in the previous chapters are mainly concerned with problems at the national level. However, the suitability of the model in connection with problems at the local and regional level should also be investigated.

In order to provide contacts with active planners and access to practical problems an active cooperation with Malmö council's statistical department has been carried on. Planners connected with the statistical department formulated the problems, which were then solved using the location model (NORLOC). The results were evaluated in a report from the statistical department (Nilsson 1972).

The problems taken up included the choice of locations for health clinics, colleges, swimming pools and old people's centers. The location model was applied with and without respect paid to the locations of existing facilities. Further a number of alternative locations were investigated.

12.1 Data for Malmö

The model requires the location of the customers to be given in square net map form. These maps are often constructed using coordinate based real estate data. The official coordinate measurement in Malmö has not yet been commenced. The allocation of a coordinate to every occupied property for the purpose of these investigations was considered to require too much time and to be too expensive. However, Malmö is divided into about 1700 statistical districts for which the population, divided into different age groups, is available on an annual basis. These areas are a block or two in size and have been allocated 5 figure reference numbers. They are therefore referred to as statistical areas at the 5 figure level. The combination of up to 10 of these areas produces a division of Malmö into 300 larger

Level	Mean	Positions				
		F1	F2	F3	F4	F5
4	2384	18900,19000				
5	2386	19000,19000				
4	1791	19900,18900	16600,19400			
5	1792	19800,18900	16600,19300			
4	1474	19900,19800	16500,19300	19400,17400		
5	1493	19800,19800	16500,19200	19400,17400		
4	1282	20200,19800	14800,18100	19400,17400	17300,19800	
5	1298	20200,19700	14800,18000	19400,17400	17200,19700	
4	1105	21800,18700	14800,18000	19400,17400	17000,19800	19600,20000
5	1124	21500,18600	14800,18000	19300,17400	17000,19800	19500,20000

Table 12.1. Mean travel distance and optimal positions for 1—5 facilities in Malmö. All values are in metres. The input data was two square net maps showing the total population at 4 and 5 figures levels. Square size: 400 metres.

statistical districts. The numbers of the smaller areas have been allocated in such a way that the larger statistical regions are defined by the first 4 figures in the reference numbers. The statistical areas have been combined to produce larger statistical areas with a fairly homogeneous character with regard to land use and other factors. A population centre of gravity was chosen manually for each statistical area at both the 4 figure and 5 figure level, thus providing coordinate based population data at two levels of detail. The statistical department in Malmö makes a population prognosis for the statistical areas at the 4 figure level. These forecasts give a predicted population for each area for the years 1975 and 1980. This information can be used as input to the location model. Before this is done a sensitivity analysis with regard to data given at the 4 figure and 5 figure levels ought to be carried out. This sensitivity analysis is similar to that described in section 7.1.3. 4 different maps were chosen for the analysis, all having an equidistance of 400 metres. These maps showed the total population at both levels of detail and the age group 16—18 at the same two levels. The leftmost column in table 12.1 shows the mean distance for the total population from home to the nearest facility. This distance is longer when the population data is given at the 5 figure level instead of the 4 figure level. The location arrangement obtained using data at the 5 figure level is therefore consistently better than that obtained using data at the 4 figure level. This conclusion is strengthened when table 12.2 is studied. This table shows mean travel distance for students, aged 16—18 years, going to 1—5 colleges in Malmö. A corresponding improvement can also be expected when one goes from data at the 5 figure level to coordinate based population data at the level of the individual property.

Level	Mean	Positions				
		F1	F2	F3	F4	F5
4	2487	18800,18800				
5	2487	18800,18800				
4	1760	20000,18800	16400,19100			
5	1752	20000,18700	16500,19000			
4	1467	20200,19800	16300,19100	19400,17400		
5	1481	20300,19500	16400,19000	19400,17400		
4	1260	20500,19700	17100,19700	15000,18000	19400,17400	
5	1266	20400,19500	17100,19600	15000,17900	19400,17400	
4	1113	19900,19800	16900,19500	14600,18000	19400,17400	22500,18900
5	1125	19800,19800	17000,19500	14900,17900	19400,17400	22400,18800

Table 12.2. Mean travel distance and optimal locations for 1—5 colleges in Malmö.

Tables 12.1 and 12.2 show the locations of 1, 2, 3, 4 and 5 facilities. In each case the locations chosen minimize the total transportation distance. The maximum difference between locations chosen in each case is 320 metres and the mean difference is 115 metres. These values are small enough to be of no practical importance since they are smaller than the equidistance. The locations chosen by the location algorithm are often already occupied by some other activity and the locations must therefore be modified to use locations which are available. Hence, this sensitivity analysis shows that one can very well use population data given at the 4 figure level when locating up to 5 facilities in Malmö.

12.2 The Malmö experiment

The choice of location for one facility serving a population, in such a way as to minimize total transportation distance, is the same as calculating the median point of the population. The population distribution is available for a number of points in time and the experiment was therefore begun by the calculation of this median point for 1960 and its movement up to 1980. The results showed a movement to the southeast during the 1960's and a continued movement to the south in the 1970's.

The population at the beginning of 1971 is available, classified into different age groups. In 1971 the median point for the age group 67 and over was to be found lying nearer to the older parts of the town than the median point for the total population in 1960. This was not the case for the 0—6 age group, the centre of which was to be found considerably nearer the newly constructed housing areas. The range around the median was also larger in this case.

12.2.1 Location of old people's centres

An old people's centre is a facility providing special services for elderly people, for example chiropodist and special gymnastics, together with rooms for such activities as weaving, pottery, TV watching and the reading of newspapers and magazines. Malmö council has built a number of these centres in new housing areas and they have exhibited very low frequencies of use. The location of old people's centres in new housing areas was done in order to encourage older people to move to these areas. This was considered desirable as the distribution of the population between age groups in the new areas tends to show a large proportion of young families. The location model was used to determine possible locations for a number of new centres. The locations chosen were in the older parts of the town, a natural result if the spatial distribution of persons in this age group is taken into account.

12.2.2 Location of colleges

There are 5 colleges (gymnasier) in Malmö at present. Their locations are shown by crosses in figure 12.1. This figure also shows the locations and umlands obtained when 5 facilities are optimally located with respect to the spatial distribution of persons in the 16—18 age group. This age group provides the largest proportion of colleges' students. An optimum location is taken to be a location which minimizes the total distance travelled to school. The square net map used for this problem had an equidistance of 400 metres and was based on the distribution of population in 1971 at the 4 figure level. One can see from figure 12.1 that the actual and optimum distributions of colleges differ considerably from one another. This is due to several reasons, the major one being that the majority of these schools are fairly old and were built before Malmö acquired the size it has today. A further reason is that there are a large number of students resident outside Malmö who go to school in Malmö. These students travel to Malmö every day by bus or train and arrive at stopping places in the centre of the city. This phenomenon can be taken into account by giving these students an origin at the stopping place in question. The third reason is that different schools provide specialized types of education such as business studies or technical courses. As a result of this a large number of pupils do not go to the school lying nearest to their home. If this phenomenon is to be taken into account then the locations for the different specialized courses must be chosen separately or the method for simultaneous hierarchical choice of location described in chapter 11 must be used.

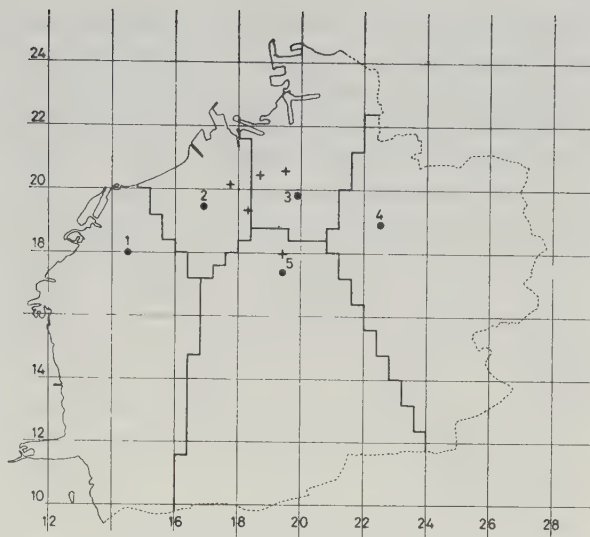


Figure 12.1. The locations and umlands for 5 optimally located colleges in Malmö. The locations of the present colleges are shown by crosses.

A study of the residential location of pupils in Lund (Nordbeck 1967) showed that the umlands of the different schools overlapped one another considerably. This is mainly due to the fact that when a pupil moves from one residence to another in the town he usually remains in the same school. Thus, he often chooses a longer journey to school instead of a changed school environment with the need to build up contacts and new friendships. As a result of this the total transportation distance incurred by the pupils when going to and from school is considerably longer than that which would occur if every pupil went to the school nearest his home. Even so, it would appear wise to locate new school facilities outside the central area of the town, this due to the reduced number of pupils resident in this part of the town.

The results provided by the location model correspond in many cases with the locations chosen by planners using other methods. The location model allows one to describe how much better or worse one alternative is than another in numerical terms, together with the degree to which it differs from an optimum arrangement. The results are therefore a valuable basis for decisions in local authority planning problems.

12.3 Locating colleges within a region

The choice of locations for schools has arisen often in this report. This is a result of the fact that the location model was constructed with the choice of locations for schools in mind. It was planned to use population data

obtained from two official data banks, the real estate data bank and the person data bank. In this way manual collection of data could have been avoided. The real estate data bank is now being built up and contains data describing 6 of Sweden's 24 counties. The authority responsible for the real estate data bank is called the Central Bureau of Real Estate Data. The person data bank is complete for every county and contains taxation information about every person, together with the name of the real estate at which he resides. The responsible authority in this case is called The National Swedish Tax Bureau (Riksskatteverket). It was planned to obtain persons' dates of birth and the code describing the property at which they reside from the person data bank. The coordinate of the property would then have been obtained from the real estate data bank. These coordinate based data were then to have been used as input to the location model.

When work with the location model was commenced it was said that the real estate data bank for Skaraborg county was practically complete. This led to the choice of the county for experimental purposes. A further reason was that the shape of the county is such that the choice of locations is not immediately obvious. The county contains a number of towns of differing size and a wide range of population densities from heavily populated areas near Lake Vänern to thinly populated forest areas in the northeast of the county.

An attempt was made to integrate chosen parts of the two data banks using the computer at the Lund Computer Centre. As a result of the fact that the codes identifying real estates differ between the two data banks it was impossible to assign coordinates to all persons in the county. About 12 000 persons, who formed about 4.5 % of the total population could not be given coordinates as a result of this. A further 50 000 persons were registered in the person data bank as being resident at a real estate having more than one separate dwelling house. The data banks do not contain information describing in which dwelling house a person resides. As a result many of these 50 000 persons received incorrect coordinates. The differences between a correct collection of coordinate based data and the results obtained from the integration of the two data banks were not randomly distributed. Instead they had a systematic character. The model is bound to be sensitive to systematic errors of this type. A manual correction of the errors would have involved considerable work and as a result Skaraborg county was abandoned as experiment area.

As a result of the difficulties described above, the long delay in completing the data bank, and day to day needs for data for planning purposes, many county authorities have begun to prepare coordinate based data

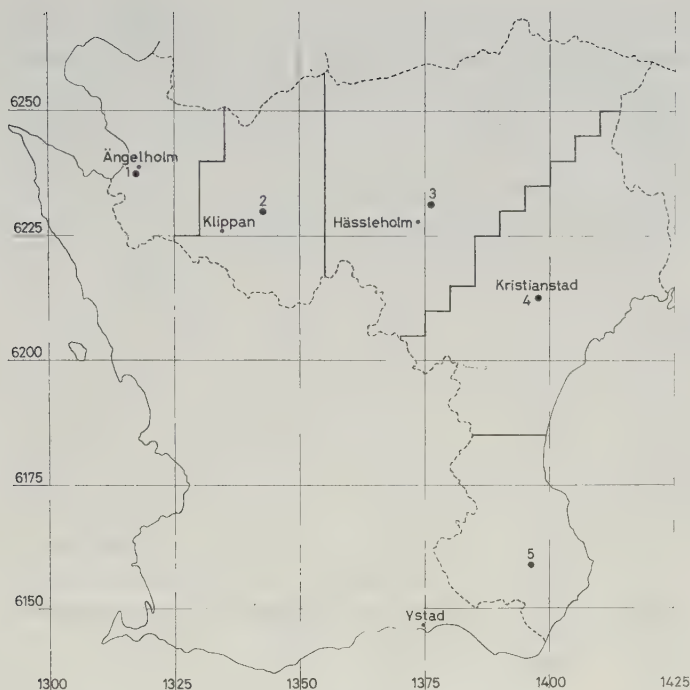


Figure 12.2. The locations and umlands for 5 optimally located colleges in Kristianstad county. The input data was a square net map, equidistance 5 km, showing that part of the county's population in 1970 aged between 16 and 18 years.

describing inhabited dwellings outside of built-up areas. All real estates within a built-up area are allocated the same coordinates, but larger built-up areas might have been divided up into statistical districts. The Swedish Central Bureau of Statistics prepares the necessary data for planning purposes for each of the statistical areas. This data includes population distribution by age, sex and occupation.

The coordinate data prepared by the individual county authorities is usually of such accuracy that a correct distribution of the population of country areas is obtained using a map with an equidistance of 1 km. Data of this type is available for Kristianstad county and as a result this county was chosen for further investigation. This county has a population approximately the same as that of the town of Malmö and Skaraborg county.

Figure 12.2 shows the locations and umlands for 5 optimally located colleges in Kristianstad county. The input data was a population square net map with an equidistance of 5 km, showing the distribution of that



Figure 12.3. At present there are colleges at Ängelholm, Klippan, Hässleholm, Kristianstad, Älmhult (Kronoberg county) and Ystad (Malmöhus county). A college is planned to start in Olofström (Blekinge county) in 1973. When facilities at these places are held fixed and one more moveable facility is to be located, the location model gives 4 local minima. The value of each minimum is given in the figure as a percentage of the lowest value. Data is the same as for figure 12.2.

part of the county's population aged between 16 and 18 years. The major difference between the result obtained in this case and the result obtained for Malmö is that in this case, except of the southern part, the optimum arrangement coincided with the actual arrangement. Ängelholm, Klippan, Hässleholm and Kristianstad each have colleges, whereas the students in the southern part of the county must go to the school in Ystad.

As a result of the fact that many students have to travel a long distance to school it is planned to build a further college in the county. In this case one should also pay attention to the fact that colleges are also to be found at Olofström in Blekinge and Älmhult in Småland. These 7 schools were given fixed locations and the location model used to determine the optimum location for a new school.

The model gave 4 different local minima. These are shown in figure 12.3 together with the relative value of each alternative when it is compared

with the most effective of the alternatives. An overall minimum is obtained when the new school is placed just to the west of Simrishamn. The students attending this school would have a mean travel distance of 11 km which should be compared with the 26 km which arises when the same pupils go to the school in Ystad. A disadvantage with the choice of Simrishamn as the new location is that the number of students attending the school in Ystad would be reduced considerably. About 1200 of the 1400 students in this area have a shorter distance to Simrishamn than to Ystad.

The next best improvement is obtained if the new school is placed near Osby. A private boarding school there already absorbs a number of students from the area. The total distance travelled by students to their respective colleges in the county would be 9 % longer if the Osby alternative is chosen instead of the Simrishamn alternative. The other two alternative locations at Perstorp and Båstad lead to an even worse overall situation. The public transport provision is fairly good in the part of Skåne in which Perstorp and Båstad lie and as a result students from these two areas can fairly easily travel to Klippan or Ängelholm.

The advantages and disadvantages of the different alternatives described here are not handled in the model, which only minimizes the total travel distance to school. The decision regarding the importance of the other factors should be made by a democratic body with political responsibilities. Hence, the location model finds its place in the investigation preliminary to the making of the decision in company with other aids.

Summary

1 The mapping system NORMAP

A standardized real estate register containing all Swedish properties is at present under preparation. This work is being carried out by the Swedish Central Bureau for Real Estate Data and should be completed before 1980. The new register contains amongst other information the geographical location of every property in coordinate form. The coordinates are, in most cases, given using the coordinate system on which the Swedish land use map is based. The new real estate register will satisfy the requirements of both planners and social researchers regarding coordinate based real estate data.

1.1 Application of coordinate information

The coordinates contained in the new register have three main uses, for identification, for sampling and for mapping. From the point of view of the social researcher the last of these is the most important. The availability of coordinate information makes possible the automatic construction of maps and diagrams showing information related to items of real estate, for example population distribution. In order for this to become possible a mapping system containing methods and computer programs is required in addition to the coordinate based property register. The development of the mapping system is therefore as important as the building up of the coordinate based data.

1.2 Computer cartography

The construction of a map involves a number of different operations such as the collection of data, the relating of the data to geographical locations, the making of decisions regarding the layout of the map, the construction of the map and finally the drawing of the map. A number of different computerized mapping systems exist, the best known of which are probably the American SYMAP and the English LINMAP. The majority of

these systems contains a well developed method for the presentation of output but, on the other hand, their data processing and map construction methods are less satisfactory. Their output is mainly based on the line printer and the map produced is usually shaded.

1.3 The NORMAP system

The NORMAP mapping system has been developed during the past 5 years at the University of Lund. The map construction and data processing sections of this system are both well developed. Also it is an integrated system which allows the processing of coordinate based data and the construction of maps on the basis of the processed data. The resulting map can be displayed in numerical form on a line printer or drawn up in its final form on a line plotter. The whole map production procedure from the collection of the data to the production of the final map has thus been made automatic.

1.4 The programming languages used in the NORMAP system

The NORMAP programs were originally written in ALGOL for the now obsolete SMIL computer at the University of Lund. In 1970 the Computer Centre of the university acquired a UNIVAC 1108 computer with a FORTRAN compiler and the programs have now been translated into FORTRAN IV.

1.5 The arrangement of the NORMAP system

A program packet of the NORMAP type can be arranged in two ways, the main program method and the subroutine method. The main program method uses a small number of large and complicated programs controlled by a correspondingly large number of parameters. A program of this type can contain so many parameters that it becomes difficult to use. The location of errors in a program of this type can also be exceedingly difficult. When using the main program method the storage space required in the computer is unnecessarily large due to the space taken up by the parts of the program not actually used at that particular instant.

The subroutine method involves the use of a large number of simple subroutines which are used in combination when carrying out complicated

tasks. This method has several advantages over the main program method. Every subroutine can be rigorously tested independent of the others. The number of parameters used at any one time is considerably reduced. Also the whole of the program is used in carrying out a task thus avoiding unnecessary waste of storage space. The addition of a subroutine to the program packet is simple, thus allowing the user to modify the system to suit special requirements. These advantages have led to the use of the subroutine method for the NORMAP system. It therefore contains only a very limited number of main programs.

1.6 NORMAP's areas of application

The subroutines contained in NORMAP can be classified into groups or blocks. An example of this type of block is the general mapping programs which construct and draw numerical and isarithmic maps. Isarithmic maps are ordinary contour maps showing, for example, height above sea level or population density. A second block of subroutines handles the mapping of population, a third fits mathematically defined surfaces to measured values using the least-square method and constructs maps showing the deviation between the calculated and actual values. A 4th block contains programs which define traffic routes and map traffic flows. The 5th block contains programs for the geographical allocation of activities, the 6th and final block contains subroutines for the analysis of directional information. These subroutines can be applied to the analysis of faults in rock formations.

1.7 Population mapping

The square net map is the simplest and therefore the cheapest type of population map. This type of map is constructed by placing a square grid over the area to be mapped and by allowing the computer to decide to which square a person should be assigned. This operation is carried out for every person in the population of the area to be mapped. The result is the number of persons assigned to each square. One of the drawbacks of the square net map is that the resulting values depend to a certain extent on where the origin of the grid has been placed. This drawback can be eliminated by allowing the squares to overlap one another; the central point in every square is then marked. These central points are usually referred to as grid points and they in their turn are used to create a new grid net which is usually triangular or rectangular in form. The value of

a square is always allocated to its grid point, i.e. its central point. The result of these operations is a numerical map which can then be processed to give an isarithmic map of the type shown in figure 1. Access to coordinate based information makes possible the construction of isarithmic maps having the same high quality as the corresponding contour maps of the landscape. As an alternative to the use of the square as a reference area a circle can be used. In this case the function mapped is the number of persons living within a distance from the grid point equal to the radius of the reference circle. This approach utilizes a constant reference area and a variable population. On the other hand, if a variable reference area and a constant population is used a new type of map is obtained, a range map. Range maps show how long one needs to "travel" from a given point in order to contain a given population.

1.8 Population potential

A range map can be used as a basis when deciding upon the siting of a service facility, as can an ordinary isarithmic map showing population. A further type of map useful for this type of problem is the potential map. In this case a value proportional to the number of visitors at every point is shown. This value is calculated on the basis of a number of assumptions. One of these is that the probability that a person will visit a certain locality is smaller the greater the distance from the person's residence to the locality in question. One should observe that the absolute value of the number of visitors at a point is not obtained; the figures refer instead to a relative frequency. On the basis of this relative frequency it is possible to determine whether the frequency of visitors at a point is larger than at some other point.

1.9. Other applications

The mapping methods developed for NORMAP can be applied in other fields. As an example of this the atomic physicists in Lund use the isarithmic program when illustrating the energy content of an atomic nucleus exposed to forces creating distortion. A further example is the use of NORMAP in conjunction with fabric diagrams. These diagrams are used by geologists when analysing faults in rock formations. The fabric diagram can also be used to determine the forces to which a rock formation has been exposed. This technique is not limited to the study of faults in rock formations which is shown by the following example.



Figure 1. An isarithmic map of Skåne with neighbouring areas of Halland, Småland and Blekinge. The map shows the distribution of rural population in 1965. The reference area is a square with sides 10 km long. The equidistance of the grid net used for interpolation is 2 km.

Some cracks appeared in a wall. These cracks showed similarities to fault structure in rock formations. At first it was believed that movement had occurred in the foundations of the building and the foundation was therefore strengthened. The cracks continued to spread even after this work had been carried out and expert help was therefore called in. A fabric analysis of the cracks was carried out and the discovery made that the cracks resulted from the roof construction. A simple alteration to the roof construction neutralized these forces and halted the cracking from spreading.

2 Location models

2.1 Central place patterns

The purest form of independent householding meant that the members of the family produced all the goods and services necessary for their existence. It was not necessary to use any service facilities outside the household itself. This pure form of independent household was abandoned at a very early stage in mankind's history and man began to purchase certain goods and services. The prerequisites for a location of these facilities at certain points were thus satisfied. The result was a network of primitive trade and service centres, which can be called central places.

There are a number of models which describe the location of central places. The best known of these is probably Christaller's central place model. In this model the assumption is made that the umlands corresponding to the different central places take the form of regular hexagons and that these hexagons cover the whole area under consideration without overlapping. The central places are nodes in a triangular net. Christaller used hexagonal umlands in order to minimize the average distance between a central place and its "customers". The points in a hexagon have a smaller mean distance to its centre than the points in a square or in a triangle. Calculation of the mean distance for hexagons and squares gives the result that the average distance for a hexagon is only 1 % less than that of the square. This difference should be considered to small to determine whether a central place has a hexagonal umland or a square or rectangular one.

2.2 Spontaneous location of service facilities

In order to simplify the control and taxation of trade and craftsmen the Swedish government introduced at an early date a regulation forbidding the carrying on of trade or crafts outside towns (*näringsstvång*). This regulation was applied up to 1864 and only after this date did a spontaneous location of trade facilities occur even in Swedish country areas. Christaller's model states that even these trade facilities should lie at the

nodes of a triangular network. Data obtained from Örkened, a parish in northeastern Skåne shows that this is not the case. On the other hand the schools in this parish are arranged in a triangular pattern, in complete accordance with Christaller's model. It appears that the locations of the schools were chosen in order to lie at the nodes of a triangular grid net.

2.3 The testing of the central place pattern in southern Sweden

A central place can be defined as an area having higher population density when compared with the surrounding area. Hence, every central place corresponds with a peak in the distribution of population density. The central place theory, amongst other things, states that these peaks are distributed in a periodic manner.

It is possible to test a collection of data for periodicity by the use of the auto-correlation method. The same method also allows the frequency of occurrence of the periodic feature and its length to be determined. An auto-correlation analysis of population density in southern Sweden indicates that if any regular pattern at all exists it is square or rectangular rather than triangular. It has thus not been possible to verify Christaller's central place pattern on the basis of data from southern Sweden.

2.4 A mathematical model of location

Location problems usually consist of the choice of optimal location for N facilities $F_1, F_2, F_3, \dots, F_N$. One assumes that the spatial distribution of the customers is given in advance, perhaps in the form of a square net map. An optimal location of facilities is taken to mean that distribution of facilities which minimizes the total cost incurred when every customer visits the facility lying nearest his dwelling. The assumption is also made that every customer has exactly the same service needs. This location model can be expressed in mathematical terms as the choice of locations which minimize the function $f(X_{F_1}, \dots, Y_{F_N})$. In this formula L refers to the number of rows in the square net map and W to the num-

$$f(X_{F_1}, \dots, Y_{F_N}) = \sum_{i=1}^L \sum_{j=1}^W P_{ij} \min_{k=1, \dots, N} (COST(D_{ij, F_k})) \quad (1)$$

ber of columns. (X_{F_k}, Y_{F_k}) are the coordinates of facility F_k . P_{ij} is the number of customers in square (i, j) of the population distribution map. $COST(D_{ij, F_k})$ is the cost incurred by a customer when travelling from

square (i,j) to facility F_k , i.e. the cost incurred when travelling distance D_{ij,F_k} . Unfortunately this function is exceedingly complicated and ordinary mathematical methods cannot be used to determine its minimum value. A heuristic iteration method must be used instead. This is done by finding a minimum value of the function $f(X_{F1}, \dots, Y_{FN})$ for the case where one of the variables $X_{F1}, Y_{F1}, \dots, Y_{FN}$ is allowed to vary while all the others are held fixed. When a minimum value is obtained, this variable is held fixed and another variable is allowed to vary until a minimum is obtained. The procedure is repeated until a stable solution is obtained. This method is known as the "naive hill-climbing method", where the problem involves the location of the highest point of a hill. This can be done by varying X , i.e. by climbing up the hill in an east-west direction. When the highest point in this direction has been reached, one continues by varying Y , i.e. by climbing along a north-south line until the highest point at this line is reached. The procedure is repeated until the highest point of the hill is reached.

2.5 Local minima

The function $f(X_{F1}, \dots, Y_{FN})$ is exceedingly complicated. It has in many cases a number of minimum points. One cannot therefore be certain whether the location algorithm has stopped at a local minimum point or whether the optimum arrangement has been reached. Due to this fact the location algorithm must be applied with great care and the results treated with restraint in order to avoid disregarding a better solution, but use of a number of different starting points for each facility reduces the risk that the algorithm will stop at a local minimum.

2.6 The transport cost function

Transport costs can be given in a table or in the form of a function. The term cost as used here should be taken in its widest meaning. It can refer to an actual cost incurred by the customer in visiting the facility but it can just as well refer to time costs etc. Formula 2 is a transport cost function of this type. $COST(D)$ is the cost incurred by a customer residing at distance D from the facility. A refers to the fixed costs incurred at ter-

$$COST(D) = A + C D^B \quad (2)$$

minals. C is a scale factor and B is an exponent. If B is equal to 1 then C can be assumed to be a price per kilometre, always assuming that D is given in kilometres.

2.7 The model's sensitivity

It has been found that the size of the square in the square net map (equidistance) can be allowed to vary within a reasonable range without any effect on the result obtained. The model is thus insensitive to reasonable variations in the size of the squares. The same is also true for variations in the exponent B of the transport cost function. The optimal location of a number of facilities was determined using the model in conjunction with 4 versions of the transportation cost function, a convex ($B > 1$), a linear ($B = 1$), a concave ($B < 1$) and a logistic version. The logistic version is characterized by a fairly small increase in transport cost $COST$, when the distance D is increased. When the value of D reaches the vicinity of a critical value E , an increase in D causes a considerable increase in $COST$. A further increase in D causes only a small scale increase in transportation cost. The 4 versions of the transport cost function produced the same location pattern for the facilities. The model is thus insensitive to reasonable variations in the value of B . It is not therefore necessary to determine transport costs to a high degree of accuracy if one has as one's main interest the choice of optimal locations for a number of facilities.

It has also been found that the location model is insensitive to reasonable errors in the input data, with the restriction that these errors have a mean value of 0. One experiment with the model involved the generation of a square net map containing errors. This map was prepared from a map that was assumed to be correct. The errors were generated using random numbers and were normally distributed with a standard deviation of 0.05 and a mean of 0. These errors were subtracted from the correct map to give a new map which was then used as input data to the location model. The result was the same location of facilities as that resulting from the use of the original correct map. The experiment was repeated with larger errors, produced by allowing the standard deviation to increase. As long as the standard deviation was under 0.50 stable solutions were obtained.

A further experiment was based on coordinate based real estate data for southern Sweden. This data was used to construct a "correct" square net map. The equidistance of this map was 5 km. The map was then used as input data to the location algorithm and the optimum location of 7 facilities determined. The next stage was the construction of a new square net map of the same area, but this time using coordinate based parish data. The equidistance of this new map was the same, i.e. 5 km. The map obtained in this way contained errors, since every resident in the parish was given the same coordinates, in this case those of the parish church.

These errors actually cancel one another out, i.e. their mean value is 0. This map was then also used as input data to the location algorithm. In both cases the locations chosen for the facilities were the same.

During the last 100 years a redistribution of the population in Sweden has occurred, the proportion of the population living in towns has increased considerably. It was considered of interest to investigate how this redistribution of population affects the location model and the results it produces. In order to carry out this study three square net maps were constructed showing the distribution of population in Sweden for the years 1855, 1917 and 1965. These maps were then used as input data to the location algorithm. The results were found to be very stable throughout the period covered. The location pattern produced is affected to an irrelevant extent by the changes in population distribution which have occurred in Sweden since the 1850's.

2.8 Regionalization

The location model can also be used to divide a region into sub-regions arranged in a hierarchical manner. One commences by dividing the area into a number of sub-regions which can then be combined to form larger regions at the next higher level in the hierarchy. When this approach is used long calculation times are incurred. An alternative is to first divide the area into a number of main regions which are then in themselves subdivided. The advantage with this technique is that the calculations can be carried out in a much shorter time and that certain requirements can be stated for the regions, such as that each sub-region should have at least a certain population.

2.9 Simultaneous location

The location model can be used in a more complicated manner in order to determine the optimal location of several different types of facility at the same time. The number of types of facility is referred to as HN . A square net map is required for every type of facility, i.e. HN square net maps are used. In addition HN transport cost functions are also required. A facility can be located individually or together with a number of other facilities. In conjunction with every start location it must be stated which facility or combination of facilities is referred to. The total transportation cost function $f(X_{F1}, \dots, Y_{FN})$ is similar to that used earlier. Due to the complicated nature of this function long calculation times are required and

the risk that the algorithm fastens at a local minimum is very large. The results should therefore be carefully checked manually. A model is always a simplified representation of reality. Therefore any model, however complicated, cannot produce results which do not require manual checking. A manual check can include those factors which are not included in the model. Different models and the use of computers can lighten the load on the planner but never replace him or his work. Against this it must be said that it is naturally incorrect to disregard these aids, nearly as incorrect as a blind belief in the results they produce.

Literature

- Abler, R., Adams, J. & Gould, P., Spatial organization. The geographer's view of the world. Englewood Cliffs 1971.
- Abrahamson, H., Geografiska synpunkter på åkerarealens struktur. — Svensk geografisk årsbok 1963.
- Améen, L., Stadsbebyggelse och domänstruktur. Svensk stadsutveckling i relation till ägoförhållanden och administrativa gränser. Lund 1964.
- Andersen, C., LLGAUSS Algol programming. — BIT 1962.
- Anscombe, F. J., Examination of residuals. — Proceeding of the fourth Berkeley symposium on mathematical statistics and probability. 1961.
- Armour, G. C. & Buffa, E. S., A heuristic algorithm and simulation approach to relative location of facilities. — Management Science 1963.
- Arpi, G. & Godlund, S., Rutkarta över Sveriges befolkning 1960. — Ymer 1963.
- Atlas över Sverige. Befolkningens fördelning 1950. Stockholm 1954.
- Bagrow, L. & Skelton, R. A., History of cartography. (Harvard University Press.). Cambridge 1964.
- Baker, L., Northwestern University computer programme library. (Department of Geography. Northwestern University). Evanston 1971.
- Bechtold, C. W., Studies in computer graphics. — Engineering Graphics 1968.
- Beckman, M. J., City hierarchies and the distributions of city size. — Economic Development and Cultural Change 1958.
- Bellman, R. E., On an optimal routing problem. — Quarterly Applied Mathematics 1958.
- Bengtsson, B. E. & Nordbeck, S., Construction of isarithms and isarithmic maps by computers. — BIT 1964.
- Bergström, L. A., Computer graphics in planning 1—8. (Department of Building Function Analysis. University of Lund). Lund 1969—1972.
- Grafisk informationsbehandling — försöksverksamhet med koordinatsatta register i Lund. (Department of Building Function Analysis. University of Lund). Lund 1971.
- Berry, B. J. L., Ribbon developments in the urban business pattern. — Annals of the Association of American Geographers 1959.
- Berry, B. J. L. & Garrison, W. L., Functional bases of the central place hierarchy. Economic Geography 1958.
- Recent developments of central place theory. — Papers and Proceedings of the Regional Science Association. Philadelphia 1958.
- Berry, B. J. L. & Marble, D. F. (eds), Spatial analysis. Englewood Cliffs 1968.
- Berry, B. J. L. & Pred, A., Central place studies. — Regional Science Research Institute Bibliography Series. No. 1. Philadelphia 1961.
- Betänkande: Underdånigt Betänkande af den för undersökningar angående Stambanornas lämpligaste sträckning i nåder tillförordnade Kommitte jämte bilagor. Stockholm 1859.

- Bickmore, D., Computers and geology. — *Geographical Magazine* 1969.
- Black, W. R., The utility of the gravity model and estimates of its parameters in commodity flow studies. — *Proceedings of the Association of American Geographers* 1971.
- Bogue, D. J., The structure of the metropolitan community. (University of Michigan). Ann Arbor 1949.
- Buffa, E. S., Armour, G. C. & Vollmann, T. E., Allocating facilities with CRAFT. — *Harvard Business Review* 1964.
- Bunge, W., Theoretical Geography. — *Lund Studies in Geography*. C 1. Lund 1966.
- Butas, L. F., A directionally oriented shortest path algorithm. — *Transportation Research* 1968.
- Butler, T., Barnens skolvägar och trafikvanor. — *Byggforskningen informerar*. Stockholm 1969.
- Bylund, E. & Weissglas, G., Glesbygdens servicenät. — *Urbaniseringen i Sverige*. SOU 1970: 14. Stockholm 1970.
- Cartwright, R. S., Calform user's manual. (Laboratory for computer graphics. Harvard University). Cambridge 1970.
- Casetti, E., Classificatory and regional analysis by discriminant iterations. (Department of Geography. Northwestern University). Evanston 1964.
- Chang, K., Adams, J. M., Rhind, D. W. & Boyd, J., Grid square mapping by computer and phototypesetter. — *Journal of the Town Planning Institute* 1971.
- Chorley, R. J. & Haggett, P., Trend surface mapping in geographical research. — *Transactions*. (Institute of British Geographers). 1965.
- *Network models in geography*. London 1969.
- Christaller, W., *Die zentralen Orte in Süddeutschland*. Jena 1933.
- Claeson, C. F., En korologisk publikanalys. — *Geografiska annaler* 1964.
- Claeson, C. F. & Stanisic, B., Koordinatkartor över befolkningen i Kronobergs län 1950 och 1960. — *Choros 10*. (Department of Social and Economic Geography. University of Gothenburg). Göteborg 1969.
- Claeson, C. F., Godlund, S. & Örtenblad, O., Population, economic structure and transportation in Ghana. — *Choros 18*. (Department of Social and Economic Geography. University of Gothenburg). Göteborg 1969.
- Clark, C., The location of industries and population. — *The Town Planning Reveiw*. 1964.
- Cliff, A. D. & Ord, J. K., Spatial autocorrelation: A review of existing and new measures with application. — *Bristol University Joint Discussion Paper in Geography and Economics*. Bristol 1969.
- Cole, A., Användande av datateknik vid planlösningsarbete. (Department of Architecture. University of Lund). Lund 1970.
- Cole, A. J., Algorithm for the production of contour maps from scattered data. — *Nature* 1968.
- Cole, J. P. & King, C. A. M., *Quantitative Geography*. London 1968.
- Connely, D. S., An experiment in contour map smoothing on the E. C. U. automated contouring system. — *Cartographic Journal* 1971.
- Cooper, L., Location-allocation problems. — *Operation Research* 1963.
- Solutions of generalized locational equilibrium models. — *Journal of Regional Science* 1967.
- An extension of the generalized Weber problem. — *Journal of Regional Science* 1968.

- Cordell, S., Digitizing contour maps. — *Datamation* 1967.
- Court, A. & Porter, P. W., The elusive point of minimum travel. — *Annals of the Association of American Geographers* 1964.
- Dahlberg, G.-B. & Ödmann, E., Stadsutveckling och planering i Sverige. Stockholm 1969.
- Dahlgren, R., Hellberg, J. & Lindberg, G., Boendeaktiviteter och omgivning. Relationernas art, innehåll, registrering och representation. — *Arbetshandling 13/71*. (Department of Building Function Analysis. University of Lund). Lund 1971.
- Dayhoff, M. O., A contour map program for X-ray cristallography. — *Communications of the ACM* 1963.
- De Dainville, F., De la profondeur à l'altide. Des origines marines de l'expression cartographique du relief terrestre par cotes et courbes de niveau. — *International Year-book of Cartography* 1962.
- Degani, A., Some computer and isodensitracer applications in geography. — *Journal of the Minnesota Academy of Science* 1969—1970.
- Automatic isoline mapping using computer, plotter and isodensitracer techniques. (University of Minnesota). 1970.
- de Geer, S., Befolkningens fördelning i Sverige år 1917. Stockholm 1919.
- De Montizon, F. A., Carte philosophique figurant la population de la France. Paris 1830.
- Dixon, W. J., BMD. Biomedical computer programs. — *University of California Publications in Automatic Computations*. No. 2. (University of California Press). Berkeley and Los Angeles 1968.
- Dueker, K. J., Organization for the retrieval of spatially distributed data: a land use data system (LUDS). — *Proceedings of the 4th annual conference on urban planning, information systems and programs*. (University of California). 1966.
- Dunn, E. S., The market potential concept and the analysis of location. — *Papers and Proceedings of the Regional Science Association* 1956.
- Erson, O. & Weissglas, G., Befolkningens fördelning på ålder och valdistrikt i W, X, Y, Z, AC och BD län. — *Glesbygdsforskningen. Rapport nr 5*. (Department of Geography. University of Umeå). Umeå 1969.
- Evans, I. S., The implementation of an automated cartography system. — *Data Processing in Biology and Geology* 1970.
- Farbey, B. A., Land, A. H. & Murchland, J., The cascade algorithm for finding the minimum distance on a graph. (Transport Network Theory Unit. London School of Economics). 1964.
- Fastighetsregistrering. Betänkande av Fastighetsregisterutredningen angående principiellt system, fastighetsbeteckningar, koordinatregistrering. — *SOU* 1966:63. Stockholm 1966.
- Bilaga: Försöksverksamhet med koordinatmetoden. (Justitiedepartementet 1967:2). Stockholm 1967.
- Koordinatregistrering av fastigheter. (Justitiedepartementet 1967:12). Stockholm 1967.
- Fisher, H. T., Reference manual for synagraphic computer mapping (SYMAP), version VI. (Laboratory for Computer Graphics. Harvard University). Cambridge 1970.
- Foot, D. H. S., The shortest route problem: ALGOL programs and discussion of computational problems in large transport network applications. — *Discussion paper 10*. (Department of Economics. University of Bristol). 1965.
- Freeman, H. & Morse, S. P., On searching a contour map for a given terrain elevation profile. — *Journal of the Franklin Institute* 1967.
- Friberg, N., Dalarnas befolkning på 1600-talet. — *Geografiska Annaler* 1953.

- Friberg, N. & Friberg, I., *Folkräkningar från 1600-talets mitt.* — Ymer 1968.
 — *En befolkningsfördelning från 1600-talets mitt.* — Ymer 1968.
 — *Sveriges äldsta fullständiga husförhörlängd.* (Department of Geography. University of Stockholm). Stockholm 1971.
- Fröberg, C. E., *Lärobok i numerisk analys.* Stockholm 1962.
- Gaits, G. M., *Thematic mapping by computers.* — *Cartographic Journal* 1969.
- Gilmore, P. C., *Optimal and suboptimal algorithms for the quadratic assignment problem.* — *Journal of the Society for Industrial and Applied Mathematics* 1962.
- Godlund, S., *Bus services, hinterlands and the location of urban settlements in Sweden, specially in Scania.* — *Studies in rural-urban interactions.* — *Lund Studies in Geography.* B 3. Lund 1951.
 — *The function and growth of bus traffic within the sphere of urban influence.* — *Lund Studies in Geography.* B 18. Lund 1956.
- Goldstein, H., *Computer mapping user study with an emphasis on planning.* (Urban Studies Center, Batelle Memorial Institute). Cleveland 1969.
- Goodchild, M. F. & Massam, B. H., *Some iterative least-cost models of spatial administrative systems in Southern Ontario.* — *Geografiska annaler.* B2. 1969.
- Gottafava, G. & Le Moli, G., *Automatic contour map.* — *Communications of the ACM* 1969.
- Gould, P., *Computers and spatial analysis: extensions of geographic research.* — *Geoforum* 1970.
- Gould, P. & Zelan, J., *The stability of autocorrelation functions from modernization surfaces (Manuscript).* (Department of Geography. The Pennsylvania State University). University Park 1971.
- Gould, P., Nordbeck, S. & Rystedt, B., *Data sensitivity and scale experiments in locating multiple facilities.* — Törnqvist et. al., *Multiple location analysis.* Lund 1971.
- Greysukh, V. L., *The possibility of studying landforms by means of digital computers.* — *Soviet Geography* 1967.
- Hagett, P., *Locational analysis in human geography.* London 1965.
- Hannerberg, D., *Täthetsisaritmer på folktäthetskartor. En studie i befolkningskartornas teori.* — *Gothia* 1937.
 — *Att studera kulturgeografi.* Stockholm 1961.
- Hanssen, B., *Österlen.* Ystad 1952.
- Harbaugh, J. W. & Merriam, D. F., *Computer applications in stratographic analysis.* New York 1968.
- Harris, C. D., *The market as a factor in the localization of industry in the United States.* — *Annals of the Association of American Geographers* 1954.
- Hedbom, O., Norling, G. & Pålsson, E., *Befolkningens fördelning i Sverige år 1960.* Stockholm 1964.
- Hedbom, O. & Norling, G., *Befolkningens fördelning i Sverige år 1965.* Stockholm 1968.
- Hellberg, J. & Lindberg, G., *Egenskaper hos fysisk och social omgivning.* — *Arbetshandling 1/69.* (Department of Building Function Analysis. University of Lund). Lund 1969.
- Hills, E. S., *Elements of structural geology.* London 1963.
- Historical statistics of Sweden.* (The Central Bureau of Statistics). Stockholm 1955.
- Horn, W., *Die Geschichte der Isarithmenkarten.* — *Petermanns Geographische Mitteilungen* 1959.
- Hsu, M.-L., *An analysis of isarithmic accuracy in relation to certain variables in the mapping process.* (University of Wisconsin). Madison 1966.

- The isopleth surface in relation to the system of data derivation. — *International Yearbook of Cartography*. 1968.
- Hsu, M.-L. & Porter, P. W., Computer mapping and geographic cartography. — *Annals of the Association of American Geographers* 1971.
- Hsu, M.-L. & Robinson, A. H., The fidelity of isopleth maps. An experimental study. — *University of Minnesota Press*. Minneapolis 1970.
- Hägerstrand, T., Statistiska primäruppgifter, flygkartering och "dataprocessing"-maskiner. Ett kombineringsprojekt. — *Svensk geografisk årsbok* 1955.
- The computer and the geographer. — *Transactions and Papers. (The Institute of British Geographers)*. 1967.
- Hägerstrand, T. & Kuklinski, A. R. (eds), Information systems for regional development — a seminar. — *Lund Studies in Geography*. B 37. Lund 1971.
- Jakobsson, A., Revision der Gemeindeeinteilung in Schweden. — *Raumforschung und Raumordnung* 1964.
- Omflyttningen i Sverige 1950—1960. Monographs published by The National Central Bureau of Statistics. Lund 1969.
- Jenks, G. F., Generalization in statistical mapping. — *Annals of the association of American Geographers* 1963.
- Jonson, J. H., Urban geography. An introductory analysis. London 1967.
- Jonston, S. R. & Roberts, J. G., Canada land inventory cartography and computer input processing of land resource data. — *La Revue de Géographie de Montréal* 1971.
- Kant, Edg., Umland studies and sector analysis. — *Studies in rural-urban interactions*. — *Lund Studies in Geography*. B 3. Lund 1951.
- Über die ersten absolute Punktkarten der Bevölkerungsverteilung. Bemerkungen zur Geschichte der thematischen Kartographie. — *Lund Studies in Geography*. B 36. Lund 1970.
- Kansky, K. J., Structure of transportation networks: relationships between network geometry and regional characteristics. — *Research Paper 84. (Department of Geography. University of Chicago)*. Chicago 1963.
- Kernighan, B. W. & Shen Lin, An efficient heuristic procedure for partitioning graphs. — *Bell System Technical Journal* 1969.
- King, L. J., Statistical analysis in geography. Englewood Cliffs 1969.
- Koeman, C. & van der Weiden, F. L. T., The application of computation and automatic drawing instruments to structural generalization. — *Cartographic Journal* 1970.
- Kolberg, D. W., Population aggregations as a continuous surface: an example of computer mapping. — *Cartographic Journal* 1970.
- Kuehn, A. A. & Hamburger, M., A heuristic program for locating warehouses. — *Management Science* 1963.
- Kuhn, H. W. & Kuenne, R. E., An efficient algorithm for the numerical solution of the generalized Weber problem in spatial economics. — *Journal of Regional Science* 1962.
- Kurtzberg, J. M., On approximation methods for the assignment problem. — *Journal of the Association for Computing Machinery* 1962.
- Larsson, I. & Fröberg, C.-E., A reconstruction of an ancient folding in pre-cambrian. — *Lund Studies in Geography*. A 6. Lund 1953.
- Lext, G., Mantalsskrivningen i Sverige före 1860. Göteborg 1968.
- Lindberg, G., Segregationsprocesser. — Lindberg, G. (ed.), *Urbana processor. Studier i social ekologi*. Lund 1971.

- Loomis, R. G. & Lorenzo, J. J., Experiments in mapping with a geo space plotter. — Urban and Regional Information Systems for Social Programs. New York 1967.
- Lösch, A., The economics of location. New Haven 1954.
- Mabougunje, A. L., Urbanization in Nigeria. Alva 1968.
- Mackay, J. R., Some problems and techniques in isopleth mapping. — Economic Geography 1951.
- Isopleth class intervals, a consideration in their selection. — The Canadian Geographer 1963.
- Malmgren, G., Avstånd, pris, sortiment. Räkneexempel kring dagligvaruhandelns utglesning. — Arbetsrapport 7/71. (Department of Building Function Analysis. University of Lund). Lund 1971.
- Marazana, F. E., On the location of supply points to minimize transport costs. — Operational Research Quarterly 1964.
- Marble, D. F., Some computer programs for geographic research. (Department of Geography. Northwestern University). Evanston 1967.
- McCarthy, H. H. & Salisbury, N. E., Visual comparison of isopleth maps as a method of determining correlations between spatially distributed phenomena. — Publication 3. (Department of Geography. State University of Iowa). 1961.
- McIntyre, D., Pollard, D. and Smith, R., Computer programs for automatic contouring. — Computer Contribution 22. (Kansas Geological Survey). Lawrence 1968.
- Merriam, D. (ed), Computer Contributions. (Kansas State Geological Survey). Lawrence 1966—1972.
- Micklander, Å. & Torstensson, I., Koordinatregister över Sveriges församlingar och tätorter. Lund 1964.
- Migneron, J.-G., Cartographie automatique et traitement des données de planification. — La Revue de Géographie de Montréal 1971.
- Mikheyeva, V. A., Application of mathematical methods and computer techniques in economic geographic research. — Soviet Geography 1965.
- Mentzer, T., Fysisk-geografisk karta öfver Skandinavien till skolornas bruk. Stockholm 1859.
- Monmonier, M. S., The production of shaded maps on the digital computer. — The Professional Geographer 1965.
- Computer mapping with the digital increment plotter. — The Professional Geographer 1968.
- On the use of digitized map sampling and measurement: an example in crop ecology. (Department of Geography. The Pennsylvania State University). University Park 1969.
- Digitized map measurement and correlation applied to an example in crop ecology. — The Geographical Review 1971.
- Morrill, R. L., Migration and the spread and growth of urban settlement. — Lund Studies in Geography. B 26. Lund 1965.
- Morrison, J. L., The effects of sampling and interpolation in isarithmic mapping. (Department of Geography. University of Wisconsin). Madison 1967.
- A link between cartographic theory and mapping practice: the nearest neighbour statistic. — The Geographical Review 1970.
- Morse, S. P., A mathematical model for the analysis of contour-line data. — Journal of the Association for Computing Machinery 1968.
- Neft, D. S., Statistical analysis for areal distributions. — Monographs Series. No. 2. (Regional Science Research Institute). Philadelphia 1967.

- Nilsson, B., Användning av koordinatmetodik för lokaliseringsberäkningar av olika serviceanläggningar i Malmö. Rapport från statistikavdelningen 3:1972. (Drätselkontorets statistikavdelning). Malmö 1972.
- Nilsson, E.-L., Skolbarnstrafik — övrig trafik. En tidsmässig konfliktsituation. — Barnens skolvägar och trafikvanor. Del 3. — Arbetsrapport 10/68. (Department of Building Function Analysis. University of Lund). Lund 1968.
- Nilsson, R., Thorén, E. & Åhlund, O., Anpassbara bostäder. Arbetshandling 12/71. (Department of Building Function Analysis. University of Lund). Lund 1971.
- Noll, A., Stereographic projections by computers. — Computers and Automation 1965.
- Norborg, K., Potentialbegreppet. Ett viktigt hjälpmedel vid lokaliseringsdiskussioner. — Svensk geografisk årsbok 1962.
- Elitlokalisering och befolkningspotential. — Svensk geografisk årsbok 1967.
- Nordbeck, S., Location of areal data for computer processing. — Lund Studies in Geography. C 2. Lund 1962.
- Computing distances in road nets. — Papers XII, Lund congress 1963. (Regional Science Association). Philadelphia 1964.
- Framställning av kartor med hjälp av siffermaskiner. Lund 1964.
- Statistiska kartor från 1800-talets mitt. — Svensk geografisk årsbok 1964.
- The law of allometric growth. — Michigan Inter-university Community of Mathematical Geographers. Ann Arbor 1965.
- Skolvägarna och trafiken. — Barnens skolvägar och trafikvanor Del 2. — Arbetsrapport 2/67. (Department of Building Function Analysis. University of Lund). Lund 1967.
- Coordinate mapping technics. — Special issue. Plan 1968.
- Koordinatsatta data och automatisk tätortsavgränsning. — Rapport från Byggforskningen 31/69. (The national Swedish institute for building research). Stockholm 1969.
- Urban allometric growth. — Geografiska annaler. B 1. 1971.
- Nordbeck, S. & Rystedt, B., Computer cartography. Point-in-polygon programs. — BIT 1967.
- Computer cartography. Range map. BIT 1969.
- Computer cartography. Shortest route programs. — Lund Studies in Geography. C 9. Lund 1969.
- Isarithmic maps and the continuity of reference interval functions. — Geografiska annaler. B 2. 1970.
- Computer Cartography. A multiple location program. — Törnqvist et. al., Multiple location analysis. Lund 1971.
- Population maps and computerized map construction. — La Revue de Géographie de Montréal 1972.
- NORMAP manual. Forthcoming.
- Nordström, O., Relationer mellan bruk och omland i östra Småland 1750—1900. Lund 1952.
- Befolkningsutveckling och arbetskraftsproblem i östra Småland 1800—1955. Lund 1957.
- Servicenäringens lokalisering och struktur i en medelstor industriort under 1900-talet. — Svensk geografisk årsbok 1965.
- Norling, G., Nya kartor över Sveriges befolkning. — Plan 1964.
- Från De Geer till Arpi-Godlund. Kommentarer till våra befolkningskartor. — Ymer 1963.
- Nugent, C. E., Vollmann, T. E. & Ruml, J., An experimental comparison of techniques for the assignment of facilities to locations. — Operation Research 1968.

- Ohlsson, B. & Nilsson, R., Anpassbara bostäder i flerfamiljshus. — Rapport från Byggnadsforskningen R22: 1970. (The national Swedish institute for building research). Stockholm 1970.
- Olsson, A. & Selander, K., Datorframställning av prickkartor. — FRIS C: 2. (Centralnämnden för fastighetsdata). Stockholm 1970.
- Olsson, G., Distance and human interaction: A review and Bibliography. — Regional Science Research Institute Bibliography Series. No 2. Philadelphia 1965.
- Olsson, R., Koordinatkarta över befolkningen i Göteborgs och Bohus' län 1960. — Choros 2. (Department of Social and Economic Geography. University of Gothenburg). Göteborg 1966.
- Ovington, J. J., Automation and map production in Australia. — Cartography 1968.
- Palmer, J. A. B., Automated Mapping. — Proceedings of 14th Australian computer conference. Adelaide 1969.
- An economical method of plotting contours. — Australian Computer Journal 1970.
- Pedersen, B., Bostadskomplement. — Arbetshandling 5/71. (Department of Building Function Analysis. University of Lund). Lund 1971.
- Pfaltz, J. L. & Rosenfeld, A., Computer representation of planar regions by their skeletons. — Communications of the ACM 1967.
- Porter, P. W., Putting the isopleth in its place. — Proceedings of the Minnesota Academy of Science. 1957—1958.
- What is the point of minimum aggregate travel. — Annals of the Association of American Geographers 1963.
- Map supplement no. 6: East Africa. Population distribution. — Annals of the Association of American Geographers 1968.
- Pålsson, E., Gymnasiers rekrytering och lokalisering. Lund 1958.
- Raisz, E., Principles of cartography. — New York 1962.
- Ravn, N. F., Populations kaart over det danske monarki 1845 og 1855. — Statistiske Tabelværk, Ny Række, Bind 12. København 1857.
- Rhind, D., Toward instant and efficient maps: the work of the experimental cartography unit. — La Revue de Géographie de Montréal 1971.
- Robertsson, L., Lösning av linjära ekvationssystem och proceduren LLGAUS. — Unpublished manuscript. Lund 1964.
- Robinson, A. H., The cartographic representation of the statistical surface. — International Yearbook of Cartography 1961.
- Mapping the correspondence of isarithmic maps. — Annals of the Association of American Geographers 1962.
- Robinson, A. H. & Caroe, L., On the analysis and comparison of statistical surfaces. — Northwestern University Studies in Geography 13. Evanston 1967.
- Robinson, A. H., Lindberg, J. B. & Brinkman, L. W., A correlation and regression analysis applied to rural farm population densities in the great plains. — Annals of the Association of American Geographers 1961.
- Robinson, A. H. & Sale, R. D., Elements of cartography. New York 1969.
- Rosenfeld, A., Picture processing by computer. — Technical Report 68—71. (Computer Science. University of Maryland). 1968.
- Rosenfeld, A. & Pfaltz, J. L., Sequential operations in digital picture processing. — Journal of the Association for Computing Machinery 1966.
- Rosing, K. E., Computer graphics. — Area 1969.
- Rosing, K. E. & Wood, P. A., Automated urban planning. — Geographical Magazine 1969.

- Rubin, J., Optimal classification into groups: An approach for solving the taxonomy problem. — *Journal of Theoretical Biology* 1967.
- Rystedt, B., Trafikkartering och trafikfördelningsmodeller. — Kompendium från konferens anordnad av Centralnämnden för fastighetsdata i maj 1969. Stockholm 1969.
- Sandblad, N. G., Skånsk stadsplanekonst och stadsarkitektur intill 1658. Lund 1949.
- Scott, A. J., Combinatorial programming, spatial analysis and planning. Birkenhead 1971.
- Shen Lin, Computer solutions to the travelling salesman problem. — *Bell System Technical Journal* 1965.
- Heuristic techniques for solving large combinatorial problems on a computer. — Lecture notes in operations research and mathematical systems theoretical approaches to non-numerical problem solving. (Springer Verlag). Berlin 1970.
- Shepard, D., A two dimensional interpolation function for irregularly spaced data. — *Proceedings of the 1968 ACM National Conference*.
- Sidenblad, G., Folkmängdstätheten och Medelegovidden i Sverige. — *Statistisk Tidskrift* 1880.
- Smailes, A. E., The geography of towns. London 1953.
- Sprunt, B., Geographics: a computer's eye view of terrain. — *Area* 1970.
- Statistical abstract of Sweden 1914—1971. (The Central Bureau of Statistics). Stockholm 1915—1972.
- Stearns, F., A method for estimating the quantitative reliability of isoline maps. — *Annals of the Association of American Geographers* 1968.
- Steinhaus, A., Über Einführung der Quadratminute und Quadratsekunde als Einheiten des geograph. Flächenmasses bei Ausmittlung der relativen Bevölkerungsdichte des flachen landes und der grösseren Städte. *Mitteil. der K.u.K. geogr. Gesellschaft zu Wien* 1863.
- Steinitz, C. (ed.), Laboratory for computer graphics and spatial analysis — selected projects. (Harvard University). Cambridge 1969.
- Stewart, J. Q., Empirical mathematical rules concerning the distribution and equilibrium of population. — *Geographical Review* 1947.
- Demographic gravitation: evidence and applications. — *Sociometry* 1948.
- Potential of population and its relationship to marketing. — *Theory in marketing*. Chicago 1950.
- A basis for social physics. — *Impact of Science on Society*. 1952.
- Svensson, H., Method for exact characterizing of denudation surfaces, especially peninsulars, as to the position in space. — *Lund Studies in Geography*. A 8. Lund 1956.
- Taaffe, E. J., Regional employment and population forecasts via relative income potential models. — *Papers and Proceedings of the Regional Science Association* 1959.
- Tarrant, J. R., Computers in geography. (University of East Anglia). Norwich 1970.
- Taylor, D. R. F., Computer mapping: a tool for the 1970's. — *La Revue de Géographie de Montréal* 1971.
- Taylor, D. R. F. & Douglas, D. H., A computer atlas of Ottawa-Hull. Ottawa 1970.
- Tobler, W. R., Automation in the preparation of thematic maps. — *Cartographic Journal* 1965.
- Of maps and matrices. — *Journal of Regional Science* 1967.
- (ed), Selected computer programs. — A Michigan Geography Publication. (Department of Geography. University of Michigan). Ann Arbor 1970.
- Tomlinson, R. F., An introduction to the geographic information system of the Canada land inventory. (Department of Forestry and Rural Development). Ottawa 1967.

- Törnqvist, G., Transport costs as a location factor for manufacturing industry. — Lund Studies in Geography. B 23. 1962.
- Studier i industrilokalisering. Stockholm 1963.
- Manufacturing goods, servicing people and processing information at multiple locations. — Törnqvist, et. al., Multiple location analysis. Lund 1971.
- Törnqvist, G., Nordbeck, S., Rystedt, B., & Gould, P., Multiple location analysis. — Lund Studies in Geography. C 12. Lund 1971.
- Ullman, E., A theory of location for cities. — American Journal of Sociology 1941.
- Vidale, M. L., A graphic solution of the transportation problem. — Journal of the Operation Research Society of America 1956.
- Villumsen, P. V., On the solution of normal equations. BIT 1965.
- Vining, R., A description of certain spatial aspects of an economic system. — Economic Development and Cultural Change 1955.
- Vitiello, D., Biggin, M. & Middleton, G., Automatic contouring at the Army Map Service. — Photogrammetric Engineering 1968.
- Voisin, R. L., Automation in private cartography. — Surveying and Mapping 1968.
- Warntz, W., Toward a geography of price. (University of Pennsylvania Press). Philadelphia 1959.
- A new map of the surface of population potentials for the United States, 1960. — Geographical Review 1964.
- Weber, A., Alfred Weber's theory of the location of industries. Translated by C. J. Friedrich. Chicago 1958.
- William-Olsson, W., Befolkningens fördelning i Sverige år 1940. Stockholm 1946.
- Williams, N. L. G., The Oxford system of automatic cartography. — Cartography 1966.
- Wolf, J. S., SYMAP: computer graphics for marketing managements. — Journal of Marketing Research 1969.
- Yeates, M., An introduction to quantitative analysis in geography. New York 1968.
- Zachariassen, J., Det Gauss. Algol programming. — BIT 1965.
- Zipf, G. K., The P_1P_2/D hypothesis: on the intercity movement of persons. — American Sociological Review 1946.
- Human behavior and the principle of least effort. Cambridge 1949.

Stig Nordbeck

Bengt Rystedt

Computer Cartography

The investigation has the character of a method study in which the authors discuss the applicability of a number of computer programs in connection with cartography, area and potential calculations and barrier situations. The second part of the investigation deals with locational models in close association with central place patterns and population distributions. The investigation is concerned with theoretical as well as methodical and verificatory problems and it can be of great value to planners of all types.

Olof Nordström

From Studentlitteratur's International Production

Andersen-Arentzen-Petersen

Asplund (Editor)

Birtwistle-Dahl-Myhrhaug-
Nygaard

Bubenko jr-Langefors-Sølvberg
(Editors)

Dopping

The Danish EDP Council

The Danish EDP Council

The Danish EDP Council

Ekman-Fröberg

Gustafsson

IAG

Johnsen

Kallin

Kurki-Suonio

Langefors

Lysegård

Naur

Nordbeck-Rystedt

Parsons & Williams

Systems Description

Management Control

SIMULA BEGIN

Computer-Aided Information Systems
Analysis and Design

Computers and Data Processing

Training Course for Programmers

Training Course for Datamaticians

Training Course for Operators

Introduction to ALGOL Programming

Optical Handwriting — Private Study
Course

Management Information Systems

Studies in Multiobjective Decision Models

Introduction to FORTRAN

Computability and Formal Languages

Theoretical Analysis of Information
Systems

Introduction to COBOL

Concise Survey of Computer Methods

Computer Cartography

Integrated Manufacturing Planning